

# An Incremental Sampling-based Algorithm for Stochastic Optimal Control

Vu Anh Huynh

Sertac Karaman

Emilio Frazzoli\*

## Abstract

In this paper, we consider a class of continuous-time, continuous-space stochastic optimal control problems. Building upon recent advances in Markov chain approximation methods and sampling-based algorithms for deterministic path planning, we propose a novel algorithm called the incremental Markov Decision Process (iMDP) to compute incrementally control policies that approximate arbitrarily well an optimal policy in terms of the expected cost. The main idea behind the algorithm is to generate a sequence of finite discretizations of the original problem through random sampling of the state space. At each iteration, the discretized problem is a Markov Decision Process that serves as an incrementally refined model of the original problem. We show that with probability one, (i) the sequence of the optimal value functions for each of the discretized problems converges uniformly to the optimal value function of the original stochastic optimal control problem, and (ii) the original optimal value function can be computed efficiently in an incremental manner using asynchronous value iterations. Thus, the proposed algorithm provides an anytime approach to the computation of optimal control policies of the continuous problem. The effectiveness of the proposed approach is demonstrated on motion planning and control problems in cluttered environments in the presence of process noise.

## 1 Introduction

Stochastic optimal control has been an active research area for several decades with many applications in diverse fields ranging from finance, management science and economics [1, 2] to biology [3] and robotics [4]. Unfortunately, general continuous-time, continuous-space stochastic optimal control problems do not admit closed-form or exact algorithmic solutions and are known to be computationally challenging [5]. Many algorithms are available to compute approximate solutions of such problems. For instance, a popular approach is based on the numerical solution of the associated Hamilton-Jacobi-Bellman PDE (see, e.g., [6, 7, 8]). Other methods approximate a continuous problem with a discrete Markov Decision Process (MDP), for which an exact solution can be computed in finite time [9, 10]. However, the complexity of these two classes of deterministic algorithms scales exponentially with the dimension of the state and control spaces, due to discretization. Remarkably, algorithms based on random (or quasi-random) sampling of the state space provide a possibility to alleviate the curse of dimensionality in the case in which the control inputs take values from a finite set, as noted in [11, 12, 5].

Algorithms based on random sampling of the state space have recently been shown to be very effective, both in theory and in practice, for computing solutions to deterministic path planning

---

\*The authors are with the Laboratory of Information and Decision Systems, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139. {vuhuynh, sertac, frazzoli}@mit.edu

problems in robotics and other disciplines. For example, the Probabilistic RoadMap (PRM) algorithm first proposed by Kavraki et al. [13] was the first practical planning algorithm that could handle high-dimensional path planning problems. Their incremental counterparts, such as RRT [14], later emerged as sampling-based algorithms suited for online applications and systems with differential constraints on the solution (e.g., dynamical systems). The RRT algorithm has been used in many applications and demonstrated on various robotic platforms [15, 16]. Recently, optimality properties of such algorithms were analyzed in [17]. In particular, it was shown that the RRT algorithm fails to converge to optimal solutions with probability one. The authors have proposed the RRT\* algorithm which guarantees almost-sure convergence to globally optimal solutions without any substantial computational overhead when compared to the RRT.

Although the RRT\* algorithm is asymptotically optimal and computationally efficient (with respect to RRT), it can not handle problems involving systems with uncertain dynamics. In this work, building upon the Markov chain approximation method [18] and the rapidly-exploring sampling technique [14], we introduce a novel algorithm called the incremental Markov Decision Process (iMDP) to approximately solve a wide class of stochastic optimal control problems. More precisely, we consider a continuous-time optimal control problem with continuous state and control spaces, full state information, and stochastic process noise. In iMDP, we iteratively construct a sequence of discrete Markov Decision Processes (MDPs) as discrete approximations to the original continuous problem, as follows. Initially, an empty MDP model is created. At each iteration, the discrete MDP is refined by adding new states sampled from the boundary as well as from the interior of the state space. Subsequently, new stochastic transitions are constructed to connect the new states to those already in the model. For the sake of efficiency, stochastic transitions are computed only when needed. Then, an anytime policy for the refined model is computed using an incremental value iteration algorithm, based on the value function of the previous model. The policy for the discrete system is finally converted to a policy for the original continuous problem. This process is iterated until convergence.

Our work is mostly related to the Stochastic Motion Roadmap (SMR) algorithm [19] and Markov chain approximation methods [18]. The SMR algorithm constructs an MDP over a sampling-based roadmap representation to maximize the probability of reaching a given goal region. However, in SMR, actions are discretized, and the algorithm does not offer any formal optimality guarantees. On the other hand, while available Markov chain approximation methods [18] provide formal optimality guarantees under very general conditions, a sequence of *a priori* discretizations of state and control spaces still impose expensive computation. The iMDP algorithm addresses this issue by sampling in the state space and sampling or discovering necessary controls.

The main contribution of this paper is a method to incrementally refine a discrete model of the original continuous problem in a way that ensures convergence to optimality while maintaining low time and space complexity. We show that with probability one, the sequence of optimal value functions induced by optimal control policies for each of the discretized problems converges uniformly to the optimal value function of the original stochastic control problem. In addition, the optimal value function of the original problem can be computed efficiently in an incremental manner using asynchronous value iterations. Thus, the proposed algorithm provides an anytime approach to the computation of optimal control policies of the continuous problem. Distributions of approximating trajectories and control processes returned from the iMDP algorithm approximate arbitrarily well distributions of optimal trajectories and optimal control processes of the original problem. Each iteration of the iMDP algorithm can be implemented with the time complexity  $O(k^\theta \log k)$  where  $0 < \theta \leq 1$  while the space complexity is  $O(k)$ , where  $k$  is the number of states in an MDP model in the algorithm which increases linearly due to the sampling strategy. Thus, the entire processing time until the algorithm stops can be implemented in  $O(k^{1+\theta} \log k)$ . Hence, the

above space and time complexities make iMDP a practical incremental algorithm. The effectiveness of the proposed approach is demonstrated on motion planning and control problems in cluttered environments in the presence of process noise.

This paper is organized as follows. In Section 2, a formal problem definition is given. The Markov chain approximation methods and the iMDP algorithm are described in Sections 3 and 4. The analysis of the iMDP algorithm is presented in Section 5. Section 6 is devoted to simulation examples and experimental results. The paper is concluded with remarks in Section 7. We provide additional notations and preliminary results as well as proofs for theorems and lemmas in Appendix.

## 2 Problem Definition

In this section, we present a generic stochastic optimal control problem. Subsequently, we discuss how the formulation extends the standard motion planning problem of reaching a goal region while avoiding collision with obstacles.

**Stochastic Dynamics** Let  $d_x$ ,  $d_u$ , and  $d_w$  be positive integers. The  $d_x$ -dimensional and  $d_u$ -dimensional Euclidean spaces are  $\mathbb{R}^{d_x}$  and  $\mathbb{R}^{d_u}$  respectively. Let  $S$  be a compact subset of  $\mathbb{R}^{d_x}$ , which is the closure of its interior  $S^\circ$  and has a smooth boundary  $\partial S$ . The state of the system at time  $t$  is  $x(t) \in S$ , which is fully observable at all times. We also define a compact subset  $U$  of  $\mathbb{R}^{d_u}$  as a control set.

Suppose that a stochastic process  $\{w(t); t \geq 0\}$  is a  $d_w$ -dimensional Brownian motion, also called a Wiener process, on some probability space  $(\Omega, \mathcal{F}, \mathcal{P})$ . Let a control process  $\{u(t); t \geq 0\}$  be a  $U$ -valued, measurable process also defined on the same probability space. We say that the control process  $u(\cdot)$  is nonanticipative with respect to the Wiener process  $w(\cdot)$  if there exists a filtration  $\{\mathcal{F}_t; t \geq 0\}$  defined on  $(\Omega, \mathcal{F}, \mathcal{P})$  such that  $u(\cdot)$  is  $\mathcal{F}_t$ -adapted, and  $w(\cdot)$  is an  $\mathcal{F}_t$ -Wiener process. In this case, we say that  $u(\cdot)$  is an admissible control inputs with respect to  $w(\cdot)$ , or the pair  $(u(\cdot), w(\cdot))$  is admissible. Let  $\mathbb{R}^{d_x \times d_w}$  denote the set of all  $d_x$  by  $d_w$  real matrices. We consider stochastic dynamical systems, also called controlled diffusions, of the form

$$dx(t) = f(x(t), u(t)) dt + F(x(t), u(t)) dw(t), \quad \forall t \geq 0 \quad (1)$$

where  $f : S \times U \rightarrow \mathbb{R}^{d_x}$  and  $F : S \times U \rightarrow \mathbb{R}^{d_x \times d_w}$  are bounded measurable and continuous functions as long as  $x(t) \in S^\circ$ . The matrix  $F(\cdot, \cdot)$  is assumed to have full rank. More precisely, a solution to the differential form given in Eq. (1) is a stochastic process  $\{x(t); t \geq 0\}$  such that  $x(t)$  equals the following stochastic integral in all sample paths:

$$x(t) = x(0) + \int_0^t f(x(\tau), u(\tau)) d\tau + \int_0^t F(x(\tau), u(\tau)) dw(\tau), \quad (2)$$

until  $x(\cdot)$  exits  $S^\circ$ , where the last term on the right hand side is the usual Itô integral (see, e.g., [20]). When the process  $x(\cdot)$  hits  $\partial S$ , the process  $x(\cdot)$  is stopped.

**Weak Existence and Weak Uniqueness of Solutions** Let  $\Gamma$  be the sample path space of admissible pairs  $(u(\cdot), w(\cdot))$ . Suppose we are given probability measures  $\Lambda$  and  $P_0$  on  $\Gamma$  and on  $S$  respectively. We say that solutions of (2) exist in the weak sense if there exists a probability space  $(\Omega, \mathcal{F}, \mathcal{P})$ , a filtration  $\{\mathcal{F}_t; t \geq 0\}$ , an  $\mathcal{F}_t$ -Wiener process  $w(\cdot)$ , an  $\mathcal{F}_t$ -adapted control process  $u(\cdot)$ , and an  $\mathcal{F}_t$ -adapted process  $x(\cdot)$  satisfying Eq. (2), such that  $\Lambda$  and  $P_0$  are the distributions of  $(u(\cdot), w(\cdot))$  and  $x(0)$  under  $\mathcal{P}$ . We call such tuple  $\{(\Omega, \mathcal{F}, \mathcal{P}), \mathcal{F}_t, w(\cdot), u(\cdot), x(\cdot)\}$  a weak sense solution of Eq. (1) [21, 18].

Assume that we are given weak sense solutions  $\{(\Omega_i, \mathcal{F}_i, \mathcal{P}_i), \mathcal{F}_{t,i}, w_i(\cdot), u_i(\cdot), x_i(\cdot)\}, i = 1, 2$ , to Eq. (1). We say solutions are weakly unique if equality of the joint distributions of  $(w_i(\cdot), u_i(\cdot), x_i(0))$  under  $\mathcal{P}_i$ ,  $i = 1, 2$ , implies the equality of the distributions  $(x_i(\cdot), w_i(\cdot), u_i(\cdot), x_i(0))$  under  $\mathcal{P}_i$ ,  $i = 1, 2$  [21, 18].

In this paper, given the boundedness of the set  $S$ , and the definition of the functions  $f$  and  $F$  in Eq. (1), we have a weak solution to Eq. (1) that is unique in the weak sense [21]. The boundedness requirement is naturally satisfied in many applications and is also needed for the implementation of the proposed numerical method. We will also handle the case in which  $f$  and  $F$  are discontinuous with extra mild technical assumptions to ensure asymptotic optimality in Section 3.

**Policy and Cost-to-go Function** A particular class of admissible controls, called *Markov controls*, depends only on the current state, i.e.,  $u(t)$  is a function only of  $x(t)$ , for all  $t \geq 0$ . It is well known that in control problems with full state information, the best Markov control performs as well as the best admissible control (see, e.g., [20, 21]). A Markov control defined on  $S$  is also called a *policy*, and is represented by the function  $\mu : S \rightarrow U$ . The set of all policies is denoted by  $\Pi$ . Define the *first exit time*  $T_\mu : \Pi \rightarrow [0, +\infty]$  under policy  $\mu$  as

$$T_\mu = \inf \{t : x(t) \notin S^\circ \text{ and Eq. (1) and } u(t) = \mu(x(t))\}.$$

Intuitively,  $T_\mu$  is the first time that the trajectory of the dynamical system given by Eq. (1) with  $u(t) = \mu(x(t))$  hits the boundary  $\partial S$  of  $S$ . By definition,  $T_\mu = +\infty$  if  $x(\cdot)$  never exits  $S^\circ$ . Clearly,  $T_\mu$  is a random variable. Then, the expected cost-to-go function under policy  $\mu$  is a mapping from  $S$  to  $\mathbb{R}$  defined as

$$J_\mu(z) = \mathbb{E} \left[ \int_0^{T_\mu} \alpha^t g(x(t), \mu(x(t))) dt + h(x(T_\mu)) \mid x(0) = z \right],$$

where  $g : S \times U \rightarrow \mathbb{R}$  and  $h : S \rightarrow \mathbb{R}$  are bounded measurable and continuous functions, called the *cost rate function* and the *terminal cost function*, respectively, and  $\alpha \in [0, 1)$  is the *discount rate*. We further assume that  $g(x, u)$  is uniformly Hölder continuous in  $x$  with exponent  $2\rho \in (0, 1]$  for all  $u \in U$ . That is, there exists some constant  $\mathcal{C} > 0$  such that

$$|g(x, u) - g(x', u)| \leq \mathcal{C} \|x - x'\|_2^{2\rho}, \quad \forall x, x' \in S.$$

We will address the discontinuity of  $g$  and  $h$  in Section 3.

The *optimal cost-to-go function*  $J^* : S \rightarrow \mathbb{R}$  is defined as  $J^*(z) = \inf_{\mu \in \Pi} J_\mu(z)$  for all  $z \in S$ . A policy  $\mu^*$  is called optimal if  $J_{\mu^*} = J^*$ . For any  $\epsilon > 0$ , a policy  $\mu$  is called an  $\epsilon$ -optimal policy if  $\|J_\mu - J^*\|_\infty \leq \epsilon$ .

In this paper, we consider the problem of computing the optimal cost-to-go function  $J^*$  and an optimal policy  $\mu^*$  if obtainable. Our approach, outlined in Section 4, approximates the optimal cost-to-go function and an optimal policy in an anytime fashion using incremental sampling-based algorithms. This sequence of approximations is guaranteed to converge uniformly to the optimal cost-to-go function and to find an  $\epsilon$ -optimal policy for an arbitrarily small non-negative  $\epsilon$ , almost surely, as the number of samples approaches infinity.

**Relationship with Standard Motion Planning** The standard motion planning problem of finding a collision-free trajectory that reaches a goal region for a deterministic dynamical system can be defined as follows (see, e.g., [17]). Let  $\mathcal{X} \subset \mathbb{R}^{d_x}$  be a compact set. Let the open sets  $\mathcal{X}_{\text{obs}}$  and  $\mathcal{X}_{\text{goal}}$  denote the obstacle region and the goal region, respectively. Define the obstacle-free space as  $\mathcal{X}_{\text{free}} := \mathcal{X} \setminus \mathcal{X}_{\text{obs}}$ . Let  $x_{\text{init}} \in \mathcal{X}_{\text{free}}$ . Consider the deterministic dynamical system

$\dot{x} = f(x(t), u(t)) dt$ , where  $f : \mathcal{X} \times U \rightarrow \mathbb{R}^{d_x}$ . The *feasible motion planning problem* is to find a measurable control input  $u : [0, T] \rightarrow U$  such that the resulting trajectory  $x(t)$  is collision free, i.e.,  $x(t) \in \mathcal{X}_{\text{free}}$  and reaches the goal region, i.e.,  $x(T) \in \mathcal{X}_{\text{goal}}$ . The *optimal motion planning problem* is to find a measurable control input  $u$  such that the resulting trajectory  $x$  solves the feasible motion planning problem with minimum trajectory cost.

The problem considered in this paper extends the classical motion planning problem with stochastic dynamics as described by Eq. (1). Given a goal set  $\mathcal{X}_{\text{goal}}$  and an obstacle set  $\mathcal{X}_{\text{obs}}$ , define  $S := \mathcal{X} \setminus (\mathcal{X}_{\text{goal}} \cup \mathcal{X}_{\text{obs}})$  and thus  $\partial\mathcal{X}_{\text{goal}} \cup \partial\mathcal{X}_{\text{obs}} \cup \partial\mathcal{X} = \partial S$ . Due to the nature of Brownian motion, under most policies, there is some non-zero probability that collision with an obstacle set will occur. However, to penalize collision with obstacles in the control design process, the cost of terminating by hitting the obstacle set, i.e.,  $h(z)$  for  $z \in \partial\mathcal{X}_{\text{obs}}$ , can be made arbitrarily high. Clearly, the higher this number is, the more conservative the resulting policy will be. Similarly, the terminal cost function on the goal set, i.e.,  $h(z)$  for  $z \in \partial\mathcal{X}_{\text{goal}}$ , can be set to a small value to encourage terminating by hitting the goal region.

### 3 Markov Chain Approximation

A discrete-state Markov decision process (MDP) is a tuple  $\mathcal{M} = (X, A, P, G, H)$  where  $X$  is a finite set of states,  $A$  is a set of actions that is possibly a continuous space,  $P(\cdot | \cdot, \cdot) : X \times X \times A \rightarrow \mathbb{R}_{\geq 0}$  is a function that denotes the transition probabilities satisfying  $\sum_{\xi' \in X} P(\xi' | \xi, v) = 1$  for all  $\xi \in X$  and all  $v \in A$ ,  $G(\cdot, \cdot) : X \times A \rightarrow \mathbb{R}$  is an immediate cost function, and  $H : X \rightarrow \mathbb{R}$  is a terminal cost function. If we start at time 0 with a state  $\xi_0 \in X$ , and at time  $i \geq 0$ , we apply an action  $v_i \in A$  at a state  $\xi_i$  to arrive at a next state  $\xi_{i+1}$  according to the transition probability function  $P$ , we have a controlled Markov chain  $\{\xi_i; i \in \mathbb{N}\}$ . The chain  $\{\xi_i; i \in \mathbb{N}\}$  due to the control sequence  $\{v_i; i \in \mathbb{N}\}$  and an initial state  $\xi_0$  will also be called the *trajectory* of  $\mathcal{M}$  under the said sequence of controls and initial state.

Given a continuous-time dynamical system as described in Eq. (1), the Markov chain approximation method approximates the continuous stochastic dynamics using a sequence of MDPs  $\{\mathcal{M}_n\}_{n=0}^{\infty}$  in which  $\mathcal{M}_n = (S_n, U, P_n, G_n, H_n)$  where  $S_n$  is a discrete subset of  $S$ , and  $U$  is the original control set. We define  $\partial S_n = \partial S \cap S_n$ . For each  $n \in \mathbb{N}$ , let  $\{\xi_i^n; i \in \mathbb{N}\}$  be a controlled Markov chain on  $\mathcal{M}_n$  until it hits  $\partial S_n$ . We associate with each state  $z$  in  $S$  a non-negative interpolation interval  $\Delta t_n(z)$ , known as a *holding time*. We define  $t_i^n = \sum_{j=0}^{i-1} \Delta t_n(\xi_j^n)$  for  $i \geq 1$  and  $t_0^n = 0$ . Let  $\Delta \xi_i^n = \xi_{i+1}^n - \xi_i^n$ . Let  $u_i^n$  denote the control used at step  $i$  for the controlled Markov chain. In addition, we define  $G_n(z, v) = g(z, v) \Delta t_n(z)$  and  $H_n(z) = h(z)$  for each  $z \in S_n$  and  $v \in U$ . Let  $\Omega_n$  be the sample space of  $\mathcal{M}_n$ . Holding times  $\Delta t_n$  and transition probabilities  $P_n$  are chosen to satisfy the *local consistency property* given by the following conditions:

1. For all  $z \in S$ ,

$$\lim_{n \rightarrow \infty} \Delta t_n(z) = 0, \quad (3)$$

2. For all  $z \in S$  and all  $v \in U$ :

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}_{P_n}[\Delta \xi_i^n | \xi_i^n = z, u_i^n = v]}{\Delta t_n(z)} = f(z, v), \quad (4)$$

$$\lim_{n \rightarrow \infty} \frac{\text{Cov}_{P_n}[\Delta \xi_i^n | \xi_i^n = z, u_i^n = v]}{\Delta t_n(z)} = F(z, v)F(z, v)^T, \quad (5)$$

$$\lim_{n \rightarrow \infty} \sup_{i \in \mathbb{N}, \omega \in \Omega_n} \|\Delta \xi_i^n\|_2 = 0. \quad (6)$$

The chain  $\{\xi_i^n; i \in \mathbb{N}\}$  is a discrete-time process. In order to approximate the continuous-time process  $x(\cdot)$  in Eq. (2), we use an *approximate continuous-time interpolation*. We define the (random) continuous-time interpolation  $\xi^n(\cdot)$  of the chain  $\{\xi_i^n; i \in \mathbb{N}\}$  and the continuous-time interpolation  $u^n(\cdot)$  of the control sequence  $\{u_i^n; i \in \mathbb{N}\}$  under the holding times function  $\Delta t_n$  as follows:  $\xi^n(\tau) = \xi_i^n$ , and  $u^n(\tau) = u_i^n$  for all  $\tau \in [t_i^n, t_{i+1}^n)$ . Let  $D^{d_x}[0, +\infty)$  denote the set of all  $\mathbb{R}^{d_x}$ -valued functions that are continuous from the left and has limits from the right. The process  $\xi^n$  can be thought of as a random mapping from  $\Omega_n$  to the function space  $D^{d_x}[0, +\infty)$ .

A control problem for the MDP  $\mathcal{M}_n$  is analogous to that defined in Section 2. Similar to previous section, a policy  $\mu_n$  is a function that maps each state  $z \in S_n$  to a control  $\mu_n(z) \in U$ . The set of all such policies is  $\Pi_n$ . Given a policy  $\mu_n$ , the (discounted) cost-to-go due to  $\mu_n$  is:

$$J_{n,\mu_n}(z) = \mathbb{E}_{P_n} \left[ \sum_{i=0}^{I_n-1} \alpha^{t_i^n} G_n(\xi_i^n, \mu_n(\xi_i^n)) + \alpha^{t_{I_n}^n} H_n(\xi_{I_n}^n) \mid \xi_0^n = z \right],$$

where  $\mathbb{E}_{P_n}$  denotes the conditional expectation under  $P_n$ , the sequence  $\{\xi_i^n; i \in \mathbb{N}\}$  is the controlled Markov chain under the policy  $\mu_n$ , and  $I_n$  is termination time defined as  $I_n = \min\{i : \xi_i^n \in \partial S_n\}$ .

The *optimal cost function*, denoted by  $J_n^*$  satisfies

$$J_n^*(z) = \inf_{\mu_n \in \Pi_n} J_{n,\mu_n}(z), \quad \forall z \in S_n. \quad (7)$$

An *optimal policy*, denoted by  $\mu_n^*$ , satisfies  $J_{n,\mu_n^*}(z) = J_n^*(z)$  for all  $z \in S_n$ . For any  $\epsilon > 0$ ,  $\mu_n$  is an  $\epsilon$ -optimal policy if  $\|J_{n,\mu_n} - J_n^*\|_\infty \leq \epsilon$ .

As stated in the following theorem, under mild technical assumptions, local consistency implies the convergence of continuous-time interpolations of the trajectories of the controlled Markov chain to the trajectories of the stochastic dynamical system described by Eq. (1).

**Theorem 1 (see Theorem 10.4.1 in [18])** *Let us assume that  $f(\cdot, \cdot)$  and  $F(\cdot, \cdot)$  are measurable, bounded and continuous. Thus, Eq. (1) has a weakly unique solution. Let  $\{\mathcal{M}_n\}_{n=0}^\infty$  be a sequence of MDPs, and  $\{\Delta t_n\}_{n=0}^\infty$  be a sequence of holding times that are locally consistent with the stochastic dynamical system described by Eq. (1). Let  $\{u_i^n; i \in \mathbb{N}\}$  be a sequence of controls defined for each  $n \in \mathbb{N}$ . For all  $n \in \mathbb{N}$ , let  $\{\xi^n(t); t \in \mathbb{R}_{\geq 0}\}$  denote the continuous-time interpolation to the chain  $\{\xi_i^n; i \in \mathbb{N}\}$  under the control sequence  $\{u_i^n; i \in \mathbb{N}\}$  starting from an initial state  $z_{\text{init}}$ , and  $\{u^n(t); t \in \mathbb{R}_{\geq 0}\}$  denote the continuous-time interpolation of  $\{u_i^n; i \in \mathbb{N}\}$ , according to the holding time  $\Delta t_n$ . Then, any subsequence of  $\{(\xi^n(\cdot), u^n(\cdot))\}_{n=0}^\infty$  has a further subsequence that converges in distribution to  $(x(\cdot), u(\cdot))$  satisfying*

$$x(t) = z_{\text{init}} + \int_0^t f(x(\tau), u(\tau)) d\tau + \int_0^t F(x(\tau), u(\tau)) dw(\tau).$$

*Under the weak uniqueness condition for solutions of Eq. (1), the sequence  $\{(\xi^n(\cdot), u^n(\cdot))\}_{n=0}^\infty$  also converges to  $(x(\cdot), u(\cdot))$ .*

Furthermore, a sequence of minimizing controls guarantees pointwise convergence of the cost function to the original optimal cost function in the following sense.

**Theorem 2 (see Theorem 10.5.2 in [18])** *Assume that  $f(\cdot, \cdot)$ ,  $F(\cdot, \cdot)$ ,  $g(\cdot, \cdot)$  and  $h(\cdot)$  are measurable, bounded and continuous. For any trajectory  $x(\cdot)$  of the system described by Eq. (1), define  $\hat{\tau}(x) := \inf\{t : x(t) \notin S^o\}$ . Let  $\{\mathcal{M}_n = (S_n, U, P_n, G_n, H_n)\}_{n=0}^\infty$  and  $\{\Delta t_n\}_{n=0}^\infty$  be locally consistent with the system described by Eq. (1).*

We suppose that the function  $\hat{\tau}(\cdot)$  is continuous (as a mapping from  $D^{d_x}[0, +\infty)$  to the compactified interval  $[0, +\infty]$ ) with probability one relative to the measure induced by any solution to Eq. (1) for an initial state  $z$ , which is satisfied when the matrix  $F(\cdot, \cdot)F(\cdot, \cdot)^T$  is nondegenerate. Then, for any  $z \in S_n$ , the following equation holds:

$$\lim_{n \rightarrow \infty} |J_n^*(z) - J^*(z)| = 0.$$

In particular, for any  $z \in S_n$ , for any sequence  $\{\epsilon_n > 0\}_{n=0}^\infty$  such that  $\lim_{n \rightarrow \infty} \epsilon_n = 0$ , and for any sequence of policies  $\{\mu_n\}_{n=0}^\infty$  such that  $\mu_n$  is an  $\epsilon_n$ -optimal policy of  $\mathcal{M}_n$ , we have:

$$\lim_{n \rightarrow \infty} |J_{n, \mu_n}(z) - J^*(z)| = 0.$$

Moreover, the sequence  $\{t_{I_n}^n; n \in \mathbb{N}\}$  converges in distribution to the termination time of the optimal control problem for the system in Eq. (1) when the system is under optimal control processes.

Under the assumption that the cost rate  $g$  is Hölder continuous [22] with exponent  $2\rho$ , the sequence of optimal value functions for approximating chains  $J_n^*$  indeed converges uniformly to  $J^*$  with a proven rate. Let us denote  $\|b\|_{S_n} = \sup_{z \in S_n} b(x)$  as the sup-norm over  $S_n$  of a function  $b$  with domain containing  $S_n$ . Let

$$\zeta_n = \max_{z \in S_n} \min_{z' \in S_n} \|z' - z\|_2 \quad (8)$$

be the dispersion of  $S_n$ .

**Theorem 3** (see Theorem 2.3 in [23] and Theorem 2.1 in [24]) *Consider an MDP sequence  $\{\mathcal{M}_n = (S_n, U, P_n, G_n, H_n)\}_{n=0}^\infty$  and holding times  $\{\Delta t_n\}_{n=0}^\infty$  that are locally consistent with the system described by Eq. (1). Let  $J_n^*$  be the optimal cost of  $\mathcal{M}_n$ . Given the assumptions on the dynamics and cost rate functions in Section 2, as  $n$  approaches  $\infty$ , we have*

$$\|J_n^* - J^*\|_{S_n} = O(\zeta_n^\rho).$$

## Discontinuity of dynamics and objective functions

We note that the above theorems continue to hold even when the functions  $f, F, g$ , and  $h$  are discontinuous. In this case, the following conditions are sufficient to use the theorems: (i) For  $r$  to be  $f, F, g$ , or  $h$ ,  $r(x, u)$  takes either the form  $r_0(x) + r_1(u)$  or  $r_0(x)r_1(u)$  where the control dependent terms are continuous and the  $x$ -dependent terms are measurable, and (ii)  $f(x, \cdot), F(x, \cdot), g(x, \cdot)$ , and  $h(x)$  are nondegenerate for each  $x$ , and the set of discontinuity in  $x$  of each function is a uniformly smooth surface of lower dimension. Furthermore, instead of uniform Hölder continuity, the cost rate  $g$  can be relaxed to be locally Hölder continuous with exponent  $2\rho$  on  $S$  (see, e.g., page 275 in [18]).

Let us remark that the controlled Markov chain differs from the stochastic dynamical systems described in Section 2 in that the former possesses a discrete state structure and evolves in a discrete time manner while the latter is a continuous model both in terms of its state space and the evolution of time. Yet, both models possess a continuous control space. It will be clear in the following discussion that the control space does not have to be discretized if a certain optimization problem can be solved numerically or via sampling.

The above theorems assert the asymptotic optimality given a sequence of *a priori* discretizations of the state space and the availability of  $\epsilon$ -optimal policies. In what follows, we describe an algorithm that incrementally computes the optimal cost-to-go function and an optimal control policy of the continuous problem.

## 4 The iMDP Algorithm

Based on Markov chain approximation results, the iMDP algorithm incrementally builds a sequence of discrete MDPs with probability transitions and cost-to-go functions that consistently approximate the original continuous counterparts. The algorithm refines the discrete models by using a number of primitive procedures to add new states into the current approximate model. Finally, the algorithm improves the quality of discrete-model policies in an iterative manner by effectively using the computations inherited from the previous iterations. Before presenting the algorithm, some primitive procedures which the algorithm relies on are presented in this section.

### 4.1 Primitive Procedures

#### 4.1.1 Sampling

The `Sample()` and `SampleBoundary()` procedures sample states independently and uniformly from the interior  $S^o$  and the boundary  $\partial S$ , respectively.

#### 4.1.2 Nearest Neighbors

Given  $z \in S$  and a set  $Y \subseteq S$  of states. For any  $k \in \mathbb{N}$ , the procedure `Nearest`( $z, Y, k$ ) returns the  $k$  nearest states  $z' \in Y$  that are closest to  $z$  in terms of the Euclidean norm.

#### 4.1.3 Time Intervals

Given a state  $z \in S$  and a number  $k \in \mathbb{N}$ , the procedure `ComputeHoldingTime`( $z, k$ ) returns a holding time computed as follows:

$$\text{ComputeHoldingTime}(z, k) = \gamma_t \left( \frac{\log k}{k} \right)^{\theta \varsigma \rho / d_x},$$

where  $\gamma_t > 0$  is a constant, and  $\varsigma, \theta$  are constants in  $(0, 1)$  and  $(0, 1]$  respectively<sup>1</sup>. The parameter  $\rho \in (0, 0.5]$  defines the Hölder continuity of the cost rate function  $g(\cdot, \cdot)$  as in Section 2.

#### 4.1.4 Transition Probabilities

Given a state  $z \in S$ , a subset  $Y \in S$ , a control  $v \in U$ , and a positive number  $\tau$  describing a holding time, the procedure `ComputeTranProb`( $z, v, \tau, Y$ ) returns (i) a finite set  $Z_{\text{near}} \subset S$  of states such that the state  $z + f(z, v)\tau$  belongs to the convex hull of  $Z_{\text{near}}$  and  $\|z' - z\|_2 = O(\tau)$  for all  $z' \neq z \in Z_{\text{near}}$ , and (ii) a function  $p$  that maps  $Z_{\text{near}}$  to a non-negative real numbers such that  $p(\cdot)$  is a probability distribution over the support  $Z_{\text{near}}$ . It is crucial to ensure that these transition probabilities result in a sequence of locally consistent chains in the algorithm.

There are several ways to construct such transition probabilities. One possible construction by solving a system of linear equations can be found in [18]. In particular, we choose  $Z_{\text{near}} = \text{Nearest}(z + f(z, v)\tau, Y, s)$  where  $s \in \mathbb{N}$  is some constant. We define the transition probabilities  $p : Z_{\text{near}} \rightarrow \mathbb{R}_{\geq 0}$  that satisfies:

- (i)  $\sum_{z' \in Z_{\text{near}}} p(z')(z' - z) = f(z, v)\tau + o(\tau)$ ,
- (ii)  $\sum_{z' \in Z_{\text{near}}} p(z')(z' - z)(z' - z)^T = F(z, v)F(z, v)^T \tau + f(z, v)f(z, v)^T \tau^2 + o(\tau)$ .

---

<sup>1</sup>Typical values of  $\varsigma$  is  $[0.999, 1)$ .



(iii)  $\sum_{z' \in Z_{\text{near}}} p(z') = 1$ .

An alternate way to compute the transition probabilities is to approximate using local Gaussian distributions. We choose  $Z_{\text{near}} = \text{Nearest}(z + f(z, v)\tau, Y, s)$  where  $s = \Theta(\log(|Y|))$ . Let  $\mathcal{N}_{\bar{m}, \sigma}(\cdot)$  denote the density of the (possibly multivariate) Gaussian distribution with mean  $\bar{m}$  and variance  $\sigma$ . Define the transition probabilities as follows:

$$p(z') = \frac{\mathcal{N}_{\bar{m}, \sigma}(z')}{\sum_{y \in Z_{\text{near}}} \mathcal{N}_{\bar{m}, \sigma}(y)},$$

where  $\bar{m} = z + f(z, v)\tau$  and  $\sigma = F(z, v)F(z, v)^T\tau$ . This expression can be evaluated easily for any fixed  $v \in U$ . As  $|Z_{\text{near}}|$  approaches infinity, the above construction satisfies the local consistency almost surely.

As we will discuss in Section 4.2, the size of the support  $Z_{\text{near}}$  affects the complexity of the iMDP algorithm. We note that solving a system of linear equations requires computing and handling a matrix of size  $(d_x^2 + d_x + 1) \times |Z_{\text{near}}|$  where  $|Z_{\text{near}}|$  is constant. When  $d_x$  and  $|Z_{\text{near}}|$  are large, the constant factor of the complexity is large. In contrast, computing local Gaussian approximation requires only  $|Z_{\text{near}}|$  evaluations. Thus, although local Gaussian approximation yields higher time complexity, this approximation is more convenient to compute.

#### 4.1.5 Backward Extension

Given  $T > 0$  and two states  $z, z' \in S$ , the procedure **ExtendBackwards**( $z, z', T$ ) returns a triple  $(x, v, \tau)$  such that (i)  $\dot{x}(t) = f(x(t), u(t))dt$  and  $u(t) = v \in U$  for all  $t \in [0, \tau]$ , (ii)  $\tau \leq T$ , (iii)  $x(t) \in S$  for all  $t \in [0, \tau]$ , (iv)  $x(\tau) = z$ , and (v)  $x(0)$  is close to  $z'$ . If no such trajectory exists, then the procedure returns failure<sup>2</sup>. We can solve for the triple  $(x, v, \tau)$  by sampling several controls  $v$  and choose the control resulting in  $x(0)$  that is closest to  $z'$ .

#### 4.1.6 Sampling and Discovering Controls

The procedure **ConstructControls**( $k, z, Y, T$ ) returns a set of  $k$  controls in  $U$ . We can uniformly sample  $k$  controls in  $U$ . Alternatively, for each state  $z' \in \text{Nearest}(z, Y, k)$ , we solve for a control  $v \in U$  such that (i)  $\dot{x}(t) = f(x(t), u(t))dt$  and  $u(t) = v \in U$  for all  $t \in [0, T]$ , (ii)  $x(t) \in S$  for all  $t \in [0, T]$ , (iii)  $x(0) = z$  and  $x(T) = z'$ .

### 4.2 Algorithm Description

The iMDP algorithm is given in Algorithm 1. The algorithm incrementally refines a sequence of (finite-state) MDPs  $\mathcal{M}_n = (S_n, U, P_n, G_n, H_n)$  and the associated holding time function  $\Delta t_n$  that consistently approximates the system in Eq. (1). In particular, given a state  $z \in S_n$  and a holding time  $\Delta t_n(z)$ , we can implicitly define the stage cost function  $G_n(z, v) = \Delta t_n(z)g(z, v)$  for all  $v \in U$  and terminal cost function  $H_n(z) = h(z)$ . We also associate with  $z \in S_n$  a cost value  $J_n(z)$ , and a control  $\mu_n(z)$ . We refer to  $J_n$  as a cost value function over  $S_n$ . In the following discussion, we describe how to construct  $S_n, P_n, J_n, \mu_n$  over iterations. We note that, in most cases, we only need to construct and access  $P_n$  on demand.

In every iteration of the main loop (Lines 4-16), we sample an additional state from the boundary of the state space  $S$ . We set  $J_n, \mu_n, \Delta t_n$  for those states at Line 5. Subsequently, we also sample a

<sup>2</sup>This procedure is used in the algorithm solely for the purpose of inheriting the “rapid exploration” property of the RRT algorithm [14, 17].

**Algorithm 1: iMDP()**

```

1   $(n, S_0, J_0, \mu_0, \Delta t_0) \leftarrow (1, \emptyset, \emptyset, \emptyset, \emptyset);$ 
2  while  $n < N$  do
3       $(S_n, J_n, \mu_n, \Delta t_n) \leftarrow (S_{n-1}, J_{n-1}, \mu_{n-1}, \Delta t_{n-1});$ 
      // Add a new state to the boundary
4       $z_s \leftarrow \text{SampleBoundary}();$ 
5       $(S_n, J_n(z_s), \mu_n(z_s), \Delta t_n(z_s)) \leftarrow (S_n \cup \{z_s\}, h(z_s), \text{null}, 0);$ 
      // Add a new state to the interior
6       $z_s \leftarrow \text{Sample}();$ 
7       $z_{\text{nearest}} \leftarrow \text{Nearest}(z_s, S_n, 1);$ 
8      if  $(x_{\text{new}}, u_{\text{new}}, \tau) \leftarrow \text{ExtendBackwards}(z_{\text{nearest}}, z_s, T_0)$  then
9           $z_{\text{new}} \leftarrow x_{\text{new}}(0);$ 
10          $\text{cost} = \tau g(z_{\text{new}}, u_{\text{new}}) + \alpha^\tau J_n(z_{\text{nearest}});$ 
11          $(S_n, J_n(z_{\text{new}}), \mu_n(z_{\text{new}}), \Delta t_n(z_{\text{new}})) \leftarrow (S_n \cup \{z_{\text{new}}\}, \text{cost}, u_{\text{new}}, \tau);$ 
        // Perform  $L_n \geq 1$  (asynchronous) value iterations
12         for  $i = 1 \rightarrow L_n$  do
            // Update  $z_{\text{new}}$  and  $K_n = \Theta(|S_n|^\theta)$  states ( $0 < \theta \leq 1, K_n < |S_n|$ )
13              $Z_{\text{update}} \leftarrow \text{Nearest}(z_{\text{new}}, S_n \setminus \partial S_n, K_n) \cup \{z_{\text{new}}\};$ 
14             for  $z \in Z_{\text{update}}$  do
15                  $\text{Update}(z, S_n, J_n, \mu_n, \Delta t_n);$ 
16      $n \leftarrow n + 1;$ 

```

state from the interior of  $S$  (Line 6) denoted as  $z_s$ . We compute the nearest state  $z_{\text{nearest}}$ , which is already in the current MDP, to the sampled state (Line 7). The algorithm computes a trajectory that reaches  $z_{\text{nearest}}$  starting at some state near  $z_s$  (Line 8) using a control signal  $u_{\text{new}}(0..\tau)$ . The new trajectory is denoted by  $x_{\text{new}} : [0, \tau] \rightarrow S$  and the starting state of the trajectory, i.e.,  $x_{\text{new}}(0)$ , is denoted by  $z_{\text{new}}$ . The new state  $z_{\text{new}}$  is added to the state set, and the cost value  $J_n(z_{\text{new}})$ , control  $\mu_n(z_{\text{new}})$ , and holding time  $\Delta t_n(z_{\text{new}})$  are initialized at Line 11.

### Update of cost value and control

The algorithm updates the cost values and controls of the finer MDP in Lines 13-15. We perform  $L_n \geq 1$  value iterations in which we update the new state  $z_{\text{new}}$  and other  $K_n = \Theta(|S_n|^\theta)$  states in the state set where  $K_n < |S_n|$ . When all states in the MDP are updated, i.e.  $K_n + 1 = |S_n|$ ,  $L_n$  value iterations are implemented in a synchronous manner. Otherwise,  $L_n$  value iterations are implemented in an asynchronous manner.

The set of states to be updated is denoted as  $Z_{\text{update}}$  (Line 13). To update a state  $z \in Z_{\text{update}}$  that is not on the boundary, in the call to the procedure **Update** (Line 15), we solve the following Bellman equation:<sup>3</sup>

$$J_n(z) = \min_{v \in U} \{G_n(z, v) + \alpha^{\Delta t_n(z)} \mathbb{E}_{P_n} [J_{n-1}(y)|z, v]\}, \quad (9)$$

<sup>3</sup>Although the argument of **Update** at Line 15 is  $J_n$ , we actually process the previous cost values  $J_{n-1}$  due to Line 3. We can implement Line 3 by simply sharing memory for  $(S_n, J_n, \mu_n, \Delta t_n)$  and  $(S_{n-1}, J_{n-1}, \mu_{n-1}, \Delta t_{n-1})$ .

<b>Algorithm 2:</b> Update( $z \in S_n, S_n, J_n, \mu_n, \Delta t_n$ )	
<pre> 1 <math>\tau \leftarrow \text{ComputeHoldingTime}(z,  S_n );</math>   // Sample or discover <math>C_n = \Theta(\log( S_n ))</math> controls 2 <math>U_n \leftarrow \text{ConstructControls}(C_n, z, S_n, \tau);</math> 3 <b>for</b> <math>v \in U_n</math> <b>do</b> 4   <math>(Z_{\text{near}}, p_n) \leftarrow \text{ComputeTranProb}(z, v, \tau, S_n);</math> 5   <math>J \leftarrow \tau g(z, v) + \alpha^\tau \sum_{y \in Z_{\text{near}}} p_n(y) J_n(y);</math> 6   <b>if</b> <math>J &lt; J_n(z)</math> <b>then</b> 7     <math>(J_n(z), \mu_n(z), \Delta t_n(z), \kappa_n(z)) \leftarrow (J, v, \tau,  S_n );</math> </pre>	

<b>Algorithm 3:</b> Policy( $z \in S, n$ )	
<pre> 1 <math>z_{\text{nearest}} \leftarrow \text{Nearest}(z, S_n, 1);</math> 2 <b>return</b> <math>\mu(z) = (\mu_n(z_{\text{nearest}}), \Delta t_n(z_{\text{nearest}}))</math> </pre>	

and set  $\mu_n(z) = v^*(z)$ , where  $v^*(z)$  is the minimizing control of the above optimization problem. There are several ways to solve Eq. (9) over the continuous control space  $U$  efficiently. If  $P_n(\cdot|z, v)$  and  $g(z, v)$  are affine functions of  $v$ , and  $U$  is convex, the above optimization has a linear objective function and a convex set of constraints. Such problems are widely studied in the literature [25]. More generally, we can uniformly sample the set of controls, called  $U_n$ , in the control space  $U$ . Hence, we can evaluate the right hand side (RHS) of Eq. (9) for each  $v \in U_n$  to find the best  $v^*$  in  $U_n$  with the smallest RHS value and thus to update  $J_n(z)$  and  $\mu_n(z)$ . When  $\lim_{n \rightarrow \infty} |U_n| = \infty$ , we can solve Eq. (9) arbitrarily well (see Theorem 8).

Thus, it is sufficient to construct the set  $U_n$  with  $\Theta(\log(|S_n|))$  controls using the procedure **ConstructControls** as described in Algorithm 2 (Line 2). The set  $Z_{\text{near}}$  and the transition probability  $P_n(\cdot|z, v)$  constructed consistently over the set  $Z_{\text{near}}$  are returned from the procedure **ComputeTranProb** for each  $v \in U_n$  (Line 4). Depending on a particular method to build  $P_n$  (i.e. solving a system of linear equations or evaluating a local Gaussian distribution), the cardinality of  $Z_{\text{near}}$  is set to a constant or increases as  $\Theta(\log(|S_n|))$ . Subsequently, the procedure chooses the best control among the constructed controls to update  $J_n(z)$  and  $\mu_n(z)$  (Line 7). We note that in Algorithm 2, before making improvement for the cost value at  $z$  by comparing new controls, we can re-evaluate the cost value with the current control  $\mu_n(z)$  over the holding time  $\Delta t_n(z)$  by adding the current control  $\mu_n(z)$  to  $U_n$ . The reason is that the current control may be still the best control compared to other controls in  $U_n$ .

## Complexity of iMDP

The time complexity per iteration of the implementation in Algorithms 1-2 is either  $O(|S_n|^\theta \log |S_n|)$  or  $O(|S_n|^\theta (\log |S_n|)^2)$ . In particular, if the procedure **ComputeTranProb** solves a set of linear equations to construct  $P_n$  such that the cardinality of  $Z_{\text{near}}$  can remain constant, the time complexity per iteration is  $O(|S_n|^\theta \log |S_n|)$  where  $\log |S_n|$  accounts for the number of processed controls, and  $|S_n|^\theta$  accounts for the number of updated states in one iteration. Otherwise, if the procedure **ComputeTranProb** uses a local Gaussian distribution to construct  $P_n$  such that the cardinality of  $Z_{\text{near}}$  increases as  $\Theta(\log |S_n|)$ , the time complexity per iteration is  $O(|S_n|^\theta (\log |S_n|)^2)$ . The processing time from the beginning until the iMDP algorithm stops after  $n$  iterations is thus either

$O(|S_n|^{1+\theta} \log |S_n|)$  or  $O(|S_n|^{1+\theta} (\log |S_n|)^2)$ . Since we only need to access locally consistent transition probability on demand, the space complexity of the iMDP algorithm is  $O(|S_n|)$ . Finally, the size of state space  $S_n$  is  $|S_n| = \Theta(n)$  due to our sampling strategy.

### 4.3 Feedback Control

As we will see in Theorems 7-8, the sequence of cost value functions  $J_n$  arbitrarily approximates the original optimal cost-to-go  $J^*$ . Therefore, we can perform a Bellman update based on the approximated cost-to-go  $J_n$  (using the stochastic continuous-time dynamics) to obtain a policy control for any  $n$ . However, we will discuss in Theorem 9 that the sequence of  $\mu_n$  also approximates arbitrarily well an optimal control policy. In other words, in the iMDP algorithm, we also incrementally construct an optimal control policy. In the following paragraph, we present an algorithm that converts a policy for a discrete system to a policy for the original continuous problem.

Given a level of approximation  $n \in \mathbb{N}$ , the control policy  $\mu_n$  generated by the iMDP algorithm is used for controlling the original system described by Eq. (1) using the procedure given in Algorithm 3. This procedure computes the state in  $\mathcal{M}_n$  that is closest to the current state of the original system and applies the control attached to this closest state over the associated holding time.

## 5 Analysis

In this section, let  $(\mathcal{M}_n = (S_n, U, P_n, G_n, H_n), \Delta t_n, J_n, \mu_n)$  denote the MDP, holding times, cost value function, and policy returned by Algorithm 1 at the end  $n$  iterations. The proofs of lemmas and theorems in this section can be found in Appendix.

For large  $n$ , states in  $S_n$  are sampled uniformly in the state space  $S$  [17]. Moreover, the dispersion of  $S_n$  shrinks with the rate  $O((\log |S_n|/|S_n|)^{1/d_x})$  as described in the next lemma.

**Lemma 4** *Recall that  $\zeta_n$  measures of the dispersion of  $S_n$  (Eq. 8). We have the following event happens with probability one:*

$$\zeta_n = O((\log |S_n|/|S_n|)^{1/d_x}).$$

The proof is based on the fact that, if we partition  $\mathbb{R}^{d_x}$  into cells of volume  $O(\log(|S_n|)/|S_n|)$ , then, almost surely, every cell contains at least an element of  $S_n$ , as  $|S_n|$  approaches infinity. The above lemma leads to the following results.

**Lemma 5** *The MDP sequence  $\{\mathcal{M}_n\}_{n=0}^\infty$  and holding times  $\{\Delta t_n\}_{n=0}^\infty$  returned by Algorithm 1 are locally consistent with the system described by Eq. (1) for large  $n$  with probability one.*

Theorem 1 and Lemma 5 together imply that the trajectories of the controlled Markov chains approximate those of the original stochastic dynamical system in Eq. (1) arbitrarily well as  $n$  approaches to infinity. Moreover, recall that  $\|\cdot\|_{S_n}$  is the sup-norm over  $S_n$ , the following theorem shows that  $J_n^*$  converges uniformly, with probability one, to the original optimal value function  $J^*$ .

**Theorem 6** *Given  $n \in \mathbb{N}$ , for all  $z \in S_n$ ,  $J_n^*(z)$  denotes the optimal value function evaluated at state  $z$  for the finite-state MDP  $\mathcal{M}_n$  returned by Algorithm 1. Then, the following event holds with probability one:*

$$\lim_{n \rightarrow \infty} \|J_n^* - J^*\|_{S_n} = 0.$$

In other words,  $J_n^*$  converges to  $J^*$  uniformly. In particular,

$$\|J_n^* - J^*\|_{S_n} = O((\log |S_n|/|S_n|)^{\rho/d_x}).$$

The proof follows immediately from Lemmas 4-5 and Theorems 2-3. The theorem suggests that we can compute  $J_n^*$  for each discrete MDP  $\mathcal{M}_n$  before sampling more states to construct  $\mathcal{M}_{n+1}$ . Indeed, in Algorithm 1, when updated states are chosen randomly as subsets of  $S_n$ , and  $L_n$  is large enough, we compute  $J_n^*$  using asynchronous value iterations [26, 27]. Subsequent theorems present stronger results.

We will prove the asymptotic optimality of the cost value  $J_n$  returned by the iMDP algorithm when  $n$  approaches infinity without directly approximating  $J_n^*$  for each  $n$ . We first consider the case when we can solve the Bellman update (Eq. 9) exactly and  $1 \leq L_n$ ,  $K_n = \Theta(|S_n|^\theta) < |S_n|$ .

**Theorem 7** *For all  $z \in S_n$ ,  $J_n(z)$  is the cost value of the state  $z$  computed by Algorithm 1 and Algorithm 2 after  $n$  iterations with  $1 \leq L_n$ , and  $K_n = \Theta(|S_n|^\theta) < |S_n|$ . Let  $J_{n,\mu_n}$  be the cost-to-go function of the returned policy  $\mu_n$  on the discrete MDP  $\mathcal{M}_n$ . If the Bellman update at Eq. 9 is solved exactly, then, the following events hold with probability one:*

- i.  $\lim_{n \rightarrow \infty} \|J_n - J_n^*\|_{S_n} = 0$ , and  $\lim_{n \rightarrow \infty} \|J_n - J^*\|_{S_n} = 0$ ,
- ii.  $\lim_{n \rightarrow \infty} |J_{n,\mu_n}(z) - J^*(z)| = 0$ ,  $\forall z \in S_n$ .

Theorem 7 enables an incremental computation of the optimal cost  $J^*$  without the need to compute  $J_n^*$  exactly before sampling more samples. Moreover, cost-to-go functions  $J_{n,\mu_n}$  induced by approximating policies  $\mu_n$  also converges pointwise to the optimal cost-to-go  $J^*$  with probability one.

When we solve the Bellman update at Eq. 9 via sampling, the following result holds.

**Theorem 8** *For all  $z \in S_n$ ,  $J_n(z)$  is the cost value of the state  $z$  computed by Algorithm 1 and Algorithm 2 after  $n$  iterations with  $1 \leq L_n$ , and  $K_n = \Theta(|S_n|^\theta) < |S_n|$ . Let  $J_{n,\mu_n}$  be the cost-to-go function of the returned policy  $\mu_n$  on the discrete MDP  $\mathcal{M}_n$ . If the Bellman update at Eq. 9 is solved via sampling such that  $\lim_{n \rightarrow \infty} |U_n| = \infty$ , then*

- i.  $\|J_n - J_n^*\|_{S_n}$  converges to 0 in probability. Thus,  $J_n$  converges uniformly to  $J^*$  in probability,
- ii.  $\lim_{n \rightarrow \infty} |J_{n,\mu_n}(z) - J^*(z)| = 0$  for all  $z \in S_n$  with probability one.

We emphasize that while the convergence of  $J_n$  to  $J^*$  is weaker than the convergence in Theorem 7, the convergence of  $J_{n,\mu_n}$  to  $J^*$  remains intact. Importantly, Theorem 1 and Theorems 7-8 together assert that starting from any initial state, trajectories and control processes provided by the iMDP algorithm approximate arbitrarily well optimal trajectories and optimal control processes of the original continuous problem. More precisely, with probability one, the induced random probability measures of approximating trajectories and approximating control processes converge weakly to the probability measures of optimal trajectories and optimal control processes of the continuous problem.

Finally, the next theorem evaluates the quality of any-time control policies returned by Algorithm 3.

**Theorem 9** *Let  $\bar{\mu}_n : S \rightarrow U$  be the interpolated policy on  $S$  of  $\mu_n : S_n \rightarrow U$  as described in Algorithm 3:*

$$\forall z \in S : \bar{\mu}_n(z) = \mu_n(y_n) \text{ where } y_n = \operatorname{argmin}_{z' \in S_n} \|z' - z\|_2.$$

Then there exists an optimal control policy  $\mu^*$  of the original problem<sup>4</sup> so that for all  $z \in S$ :

$$\lim_{n \rightarrow \infty} \bar{\mu}_n(z) = \mu^*(z) \text{ w.p.1,}$$

if  $\mu^*$  is continuous at  $z$ .

## 6 Experiments

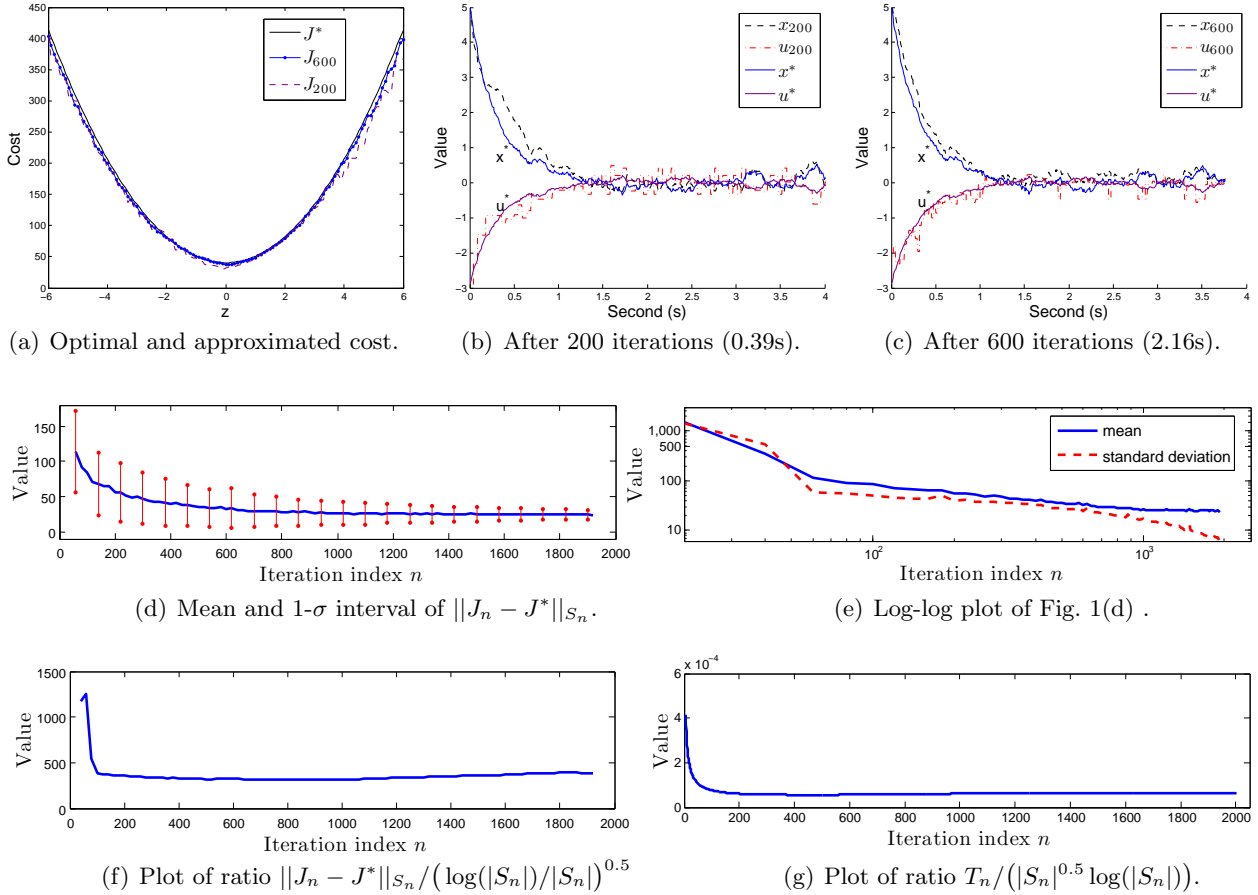


Figure 1: Results of iMDP on a stochastic LQR problem. Figure 1(a) shows the convergence of approximated cost-to-go to the optimal analytical cost-to-go over iterations. Anytime solutions are compared to the analytical optimal solution after 200 and 600 iterations in Figs. 1(b)-1(c). Mean and 1- $\sigma$  interval of the error  $\|J_n - J^*\|_{S_n}$  are shown in 1(d) using 50 trials. The corresponding mean and standard deviation of the error  $\|J_n - J^*\|_{S_n}$  are depicted on a log-log plot in Fig. 1(e). In Fig. 1(f), we plot the ratio of  $\|J_n - J^*\|_{S_n}$  to  $(\log(|S_n|)/|S_n|)^{0.5}$  to show the convergence rate of  $J_n$  to  $J^*$ . Figure 1(g) shows the ratio of running time per iteration  $T_n$  to  $|S_n|^{0.5} \log(|S_n|)$ . Ratios in Figs. 1(f)-1(g) are averaged over 50 trials.

We used a computer with a 2.0-GHz Intel Core 2 Duo T6400 processor and 4 GB of RAM to run experiments. In the first experiment, we investigated the convergence of the iMDP algorithm on a stochastic LQR problem:  $\inf \mathbb{E} \left[ \int_0^T 0.95^t \{3.5x(t)^2 + 200u(t)^2\} dt + h(x(\tau)) \right]$  such that  $dx =$

<sup>4</sup>Otherwise, an optimal relaxed control policy  $m^*$  exists [18], and  $\bar{\mu}_n$  approximates  $m^*$  arbitrarily well.

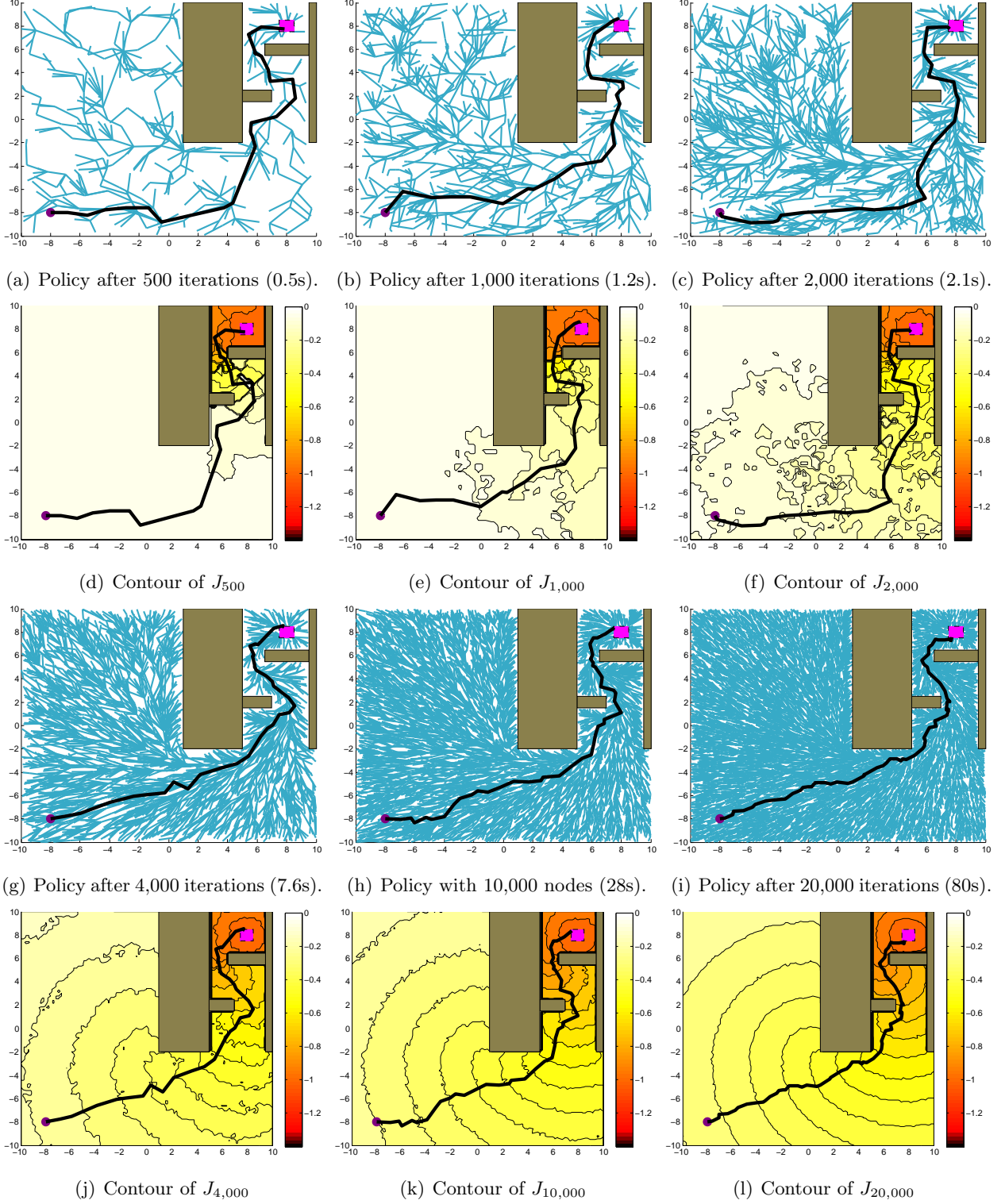


Figure 2: A system with stochastic single integrator dynamics in a cluttered environment. With appropriate cost structure assigned to the goal and obstacle regions, the system reaches the goal in the upper right corner and avoids obstacles. The standard deviation of noise in x and y directions is 0.26. The maximum velocity is one. Anytime control policies and corresponding contours of approximated cost-to-go as shown in Figs. 2(a)-2(l) indicate that iMDP quickly explores the state space and refines control policies over time.

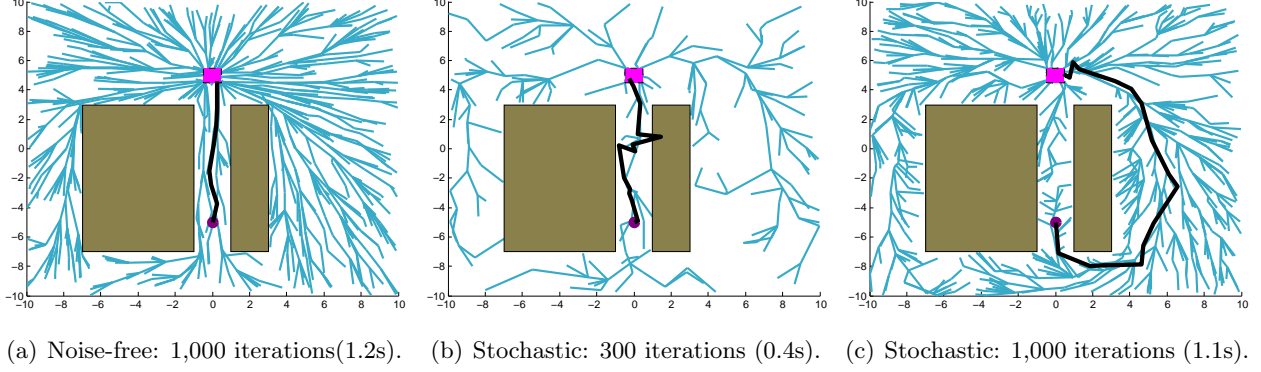


Figure 3: Performance against different process noise magnitude. The system starts from (0,-5) to reach the goal. In Fig. 3(a), the environment is noise-free. In Figs. 3(b)-3(c), standard deviation of noise in x and y directions is 0.37. In the latter, the system first discovers an unsafe route that is prone to collisions and discovers a safer route after a few seconds. (In Fig. 3(b), we temporarily let the system continue even after collision to observe the entire trajectory.)

$(3x+11u)dt + \sqrt{0.2}dw$  on the state space  $S = [-6, 6]$  where  $\tau$  is the first hitting time to the boundary  $\partial S = \{-6, 6\}$ , and  $h(z) = 414.55$  for  $z \in \partial S$  and 0 otherwise. The optimal cost-to-go from  $x(0) = z$  is  $10.39z^2 + 40.51$ , and the optimal control policy is  $u(t) = -0.5714x(t)$ . Since the cost-rate function is bounded on  $S$  and Hölder continuous with exponent 1.0, we use  $\rho = 0.5$ . In addition, we choose  $\theta = 0.5$ , and  $\varsigma = 0.99$  in the procedure `ComputeHoldingTime`. We used the procedure `Update` as presented in Algorithm 2 with  $\log(n)$  sampled controls and transition probabilities having constant support size. Figures 1(a)-1(c) show the convergence of approximated cost-to-go, anytime controls and trajectory to the optimal analytical counterparts over iterations. We observe that in Fig. 1(d), both the mean and variance of cost-to-go error decreases quickly to zero. The log-log plot in Fig. 1(e) clearly indicates that both mean and standard deviation of the error  $\|J_n - J^*\|_{S_n}$  continue to decrease. This observation is consistent with Theorems 7-8. Moreover, Fig. 1(f) shows the ratio of  $\|J_n - J^*\|_{S_n}$  to  $(\log(|S_n|)/|S_n|)^{0.5}$  indicating the convergence rate of  $J_n$  to  $J^*$ , which agrees with Theorem 6. Finally, Fig. 1(g) plots the ratio of running time per iteration  $T_n$  to  $|S_n|^{0.5} \log(|S_n|)$  asserting that the time complexity per iteration is  $O(|S_n|^{0.5} \log(|S_n|))$ .

In the second experiment, we controlled a system with stochastic single integrator dynamics to a goal region with free ending time in a cluttered environment. The cost objective function is discounted with  $\alpha = 0.95$ . The system pays zero cost for each action it takes and pays a cost of -1 when reaching the goal region  $\mathcal{X}_{goal}$ . The maximum velocity of the system is one. The system stops when it collides with obstacles. We show how the system reaches the goal in the upper right corner and avoids obstacles with different anytime controls. Anytime control policies after up-to 2,000 iterations in Figs. 2(a)-2(c), which were obtained within 2.1 seconds, indicate that iMDP quickly explores the state space and refines control policies over time. Corresponding contours of cost value functions are shown in Figs. 2(d)-2(f) further illustrate the refinement and convergence of cost value functions to the original optimal cost-to-go over time. We observe that the performance is suitable for real-time control. Furthermore, anytime control policies and cost value functions after up-to 20,000 iterations are shown in Figs. 2(g)-2(i) and Figs. 2(j)-2(l) respectively. We note that the control policies seem to converge faster than cost value functions over iterations. The phenomenon is due to the fact that cost value functions  $J_n$  are the estimates of the optimal cost-to-go  $J^*$ . Thus, when  $J_n(z) - J^*(z)$  is constant for all  $z \in S_n$ , updated controls after a Bellman update are close to their optimal values. Thus, the phenomenon favors the use of the iMDP algorithm in real-time



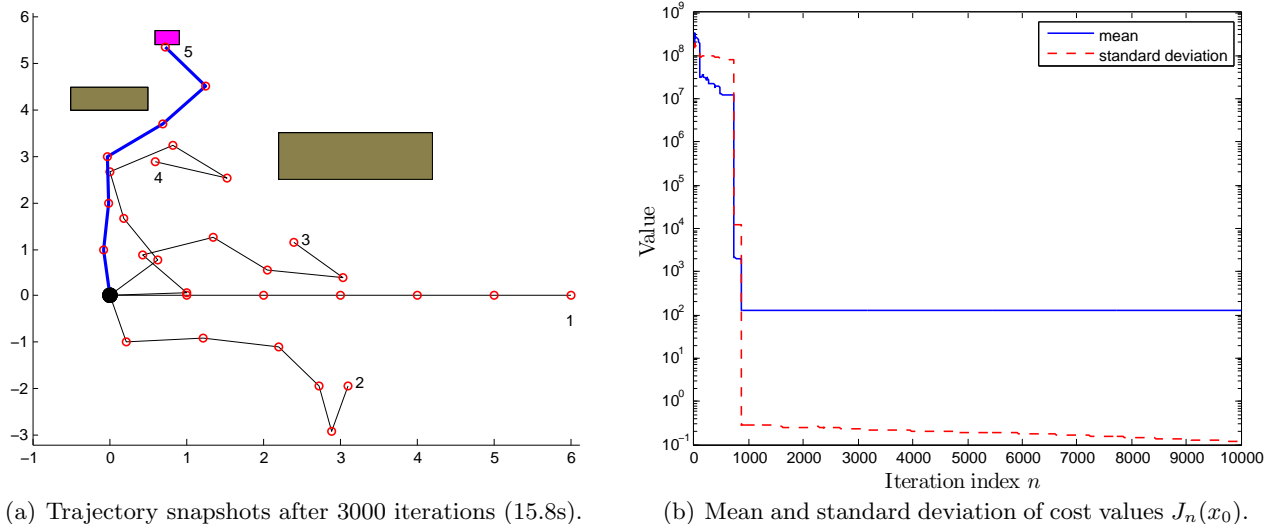


Figure 4: Results of a 6D manipulator example. The system is modeled as a single integrator with states representing angles between segments and the horizontal line. Control magnitude is bounded by 0.3. The standard deviation of noise at each joint is 0.032 rad. In Fig. 4(a), the manipulator is controlled to reach a goal with the final upright position. In Fig. 4(b), the mean and standard deviation of the computed cost values for the initial position are plotted using 50 trials.

applications where only a small number of iterations are executed.

In the third experiment, we tested the effect of process noise magnitude on the solution trajectories. In Figs. 3(a)-3(c), the system wants to arrive at a goal area either by passing through a narrow corridor or detouring around the two blocks. In Fig. 3(a), when the dynamics is noise-free (by setting a small diffusion matrix), the iMDP algorithm quickly determines to follow a narrow corridor. In contrast, when the environment affects the dynamics of the system (Figs. 3(b)-3(c)), the iMDP algorithm decides to detour to have a safer route. This experiment demonstrates the benefit of iMDP in handling process noise compared to RRT-like algorithms [14, 17]. We emphasize that although iMDP spends slightly more time on computation per iteration, iMDP provides feedback policies rather than open-loop policies; thus, re-planning is not crucial in iMDP.

In the fourth experiment, we examined the performance of the iMDP algorithm for high dimensional systems such as a manipulator with six degrees of freedom. The manipulator is modeled as a single integrator where states represents angles between segments and the horizontal line. The maximum control magnitude for all joints is 0.3. The standard deviation of noise at each joint is 0.032 rad. The manipulator is controlled to reach a goal with the final upright position in minimum time. In Fig. 4(a), we show a resulting trajectory after 3000 iterations computed in 15.8 seconds. In addition, we show the mean and standard deviation of the computed cost values for the initial position using 50 trials in Fig. 4(b). As shown in the plots, the solution converges quickly after about 1000 iterations. These results highlight the suitability of the iMDP algorithm to compute feedback policies for complex high dimensional systems in stochastic environments.

## 7 Conclusions

We have introduced and analyzed the incremental sampling-based iMDP algorithm for stochastic optimal control. The algorithm natively handles continuous time, continuous state space as well as continuous control space. The main idea is to consistently approximate underlying continuous

problems by discrete structures in an incremental manner. In particular, we incrementally build discrete MDPs by sampling and extending states in the state space. The iMDP algorithm refines the quality of anytime control policies from discrete MDPs in terms of expected costs over iterations and ensures almost sure convergence to an optimal continuous control policy. The iMDP algorithm can be implemented such that its time complexity per iteration grows as  $O(k^\theta \log k)$  with  $0 < \theta \leq 1$  leading to the total processing time  $O(k^{1+\theta} \log k)$ , where  $k$  is the number of states in MDPs which increases linearly over iterations. Together with linear space complexity, iMDP is a practical incremental algorithm. The enabling technical ideas lie in novel methods to compute Bellman updates.

Further extension of the work is broad. In the future, we would like to study the effect of biased-sampling techniques on the performance of iMDP. The algorithm is also highly parallelizable, and efficient parallel versions of the iMDP algorithm are left for future study. Remarkably, Markov chain approximation methods are also tools to handle deterministic control and non-linear filtering problems. Thus, applications of the iMDP algorithm can be extended to classical path planning with deterministic dynamics. We emphasize that the iMDP algorithm would remove the necessity for exact point-to-point steering of RRT-like algorithms in path planning applications. In addition, we plan to investigate incremental sampling-based algorithms for online smoothing and estimation in the presence of sensor noise. The combination of incremental sampling-based algorithms for control and estimation will provide insights into addressing stochastic optimal control problems with imperfect state information, known as Partially Observable Markov Decision Processes (POMDPs). Although POMDPs are fundamentally more challenging than the problem that is studied in this paper, our approach differentiates itself from existing sampling-based POMDP solvers (see, e.g., [28, 29]) with its incremental nature and computationally-efficient search. Hence, the research presented in this paper opens a new alley to handle POMDPs in our future work.

## ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation, grant CNS-1016213. V. A. Huynh gratefully thanks the Arthur Gelb Foundation for supporting him during this work.

## References

- [1] W. H. Fleming and J. L. Stein, “Stochastic optimal control, international finance and debt,” *Journal of Banking and Finance*, vol. 28, pp. 979–996, 2004.
- [2] S. P. Sethi and G. L. Thompson, *Optimal Control Theory: Applications to Management Science and Economics*, 2nd ed. Springer, 2006.
- [3] E. Todorov, “Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system,” *Neural Computation*, vol. 17, pp. 1084–1108, 2005.
- [4] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*, 2001.
- [5] V. D. Blondel and J. N. Tsitsiklis, “A survey of computational complexity results in systems and control,” *Automatica*, vol. 36, no. 9, pp. 1249–1274, 2000.
- [6] L. Grne, “An adaptive grid scheme for the discrete hamilton-jacobi-bellman equation,” *Numerische Mathematik*, vol. 75, pp. 319–337, 1997.

- [7] S. Wang, L. S. Jennings, and K. L. Teo, "Numerical solution of hamilton-jacobi-bellman equations by an upwind finite volume method," *J. of Global Optimization*, vol. 27, pp. 177–192, November 2003.
- [8] M. Boulbrachene and B. Chentouf, "The finite element approximation of hamilton-jacobi-bellman equations: the noncoercive case," *Applied Mathematics and Computation*, vol. 158, no. 2, pp. 585–592, 2004.
- [9] C. Chow and J. Tsitsiklis, "An optimal one-way multigrid algorithm for discrete-time stochastic control," *IEEE Transactions on Automatic Control*, vol. AC-36, pp. 898–914, 1991.
- [10] R. Munos, A. Moore, and S. Singh, "Variable resolution discretization in optimal control," in *Machine Learning*, 2001, pp. 291–323.
- [11] J. Rust, "Using Randomization to Break the Curse of Dimensionality,," *Econometrica*, vol. 56, no. 3, May 1997.
- [12] —, "A comparison of policy iteration methods for solving continuous-state, infinite-horizon markovian decision problems using random, quasi-random, and deterministic discretizations," 1997.
- [13] L. E. Kavraki, P. Svestka, L. E. K. P. Vestka, J. claude Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Transactions on Robotics and Automation*, vol. 12, pp. 566–580, 1996.
- [14] S. M. Lavalle, "Rapidly-exploring random trees: A new tool for path planning," Tech. Rep., 1998.
- [15] J. Kim and J. P. Ostrowski, "Motion planning of aerial robot using rapidly-exploring random trees with dynamic constraints," in *ICRA*, 2003, pp. 2200–2205.
- [16] Y. Kuwata, J. Teo, G. Fiore, S. Karaman, E. Frazzoli, and J. How, "Real-time motion planning with applications to autonomous urban driving," *IEEE Trans. on Control Systems Technologies*, vol. 17, no. 5, pp. 1105–1118, 2009.
- [17] Karaman and Frazzoli, "Sampling-based algorithms for optimal motion planning," *International Journal of Robotics Research*, vol. 30, no. 7, pp. 846–894, June 2011.
- [18] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time (Stochastic Modelling and Applied Probability)*. Springer, Dec. 2000.
- [19] R. Alterovitz, T. Simon, and K. Goldberg, "The stochastic motion roadmap: A sampling framework for planning with markov motion uncertainty," in *in Robotics: Science and Systems III (Proc. RSS 2007)*. MIT Press, 2008, pp. 246–253.
- [20] B. Oksendal, *Stochastic differential equations (3rd ed.): an introduction with applications*. New York, NY, USA: Springer-Verlag New York, Inc., 1992.
- [21] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus (Graduate Texts in Mathematics)*, 2nd ed. Springer, Aug. 1991.
- [22] L. C. Evans, *Partial Differential Equations (Graduate Studies in Mathematics, V. 19) GSM/19*. American Mathematical Society, Jun. 1998.

- [23] J. L. Menaldi, “Some estimates for finite difference approximations,” *SIAM J. on Control and Optimization*, vol. 27, pp. 579–607, 1989.
- [24] P. Dupuis and M. R. James, “Rates of convergence for approximation schemes in optimal control,” *SIAM J. Control Optim.*, vol. 36, pp. 719–741, March 1998.
- [25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Mar. 2004.
- [26] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Two Volume Set*, 2nd ed. Athena Scientific, 2001.
- [27] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1989.
- [28] H. Kurniawati, D. Hsu, and W. Lee, “SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces,” in *Proc. Robotics: Science and Systems*, 2008.
- [29] S. Prentice and N. Roy, “The belief roadmap: Efficient planning in linear pomdps by factoring the covariance,” in *Proceedings of the 13th International Symposium of Robotics Research (ISRR)*, Hiroshima, Japan, November 2007.
- [30] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford University Press, USA, Aug. 2001.

## Appendix

### A Notations and Preliminaries

We denote  $\mathbb{N}$  as a set of natural numbers and  $\mathbb{R}$  as a set of real numbers. A sequence on a set  $X$  is a mapping from  $\mathbb{N}$  to  $X$ , denoted as  $\{x_n\}_{n=0}^{\infty}$ , where  $x_n \in X$  for each  $n \in \mathbb{N}$ . Given a metric space  $X$  endowed with a metric  $d$ , a sequence  $\{x_n\}_{n=0}^{\infty} \subset X$  is said to converge if there is a point  $x \in X$ , denoted as  $\lim_{n \rightarrow \infty} x_n$ , with the following property: For every  $\epsilon > 0$ , there is an integer  $N$  such that  $n \geq N$  implies that  $d(x_n, x) < \epsilon$ . On the one hand, a sequence of functions  $\{f_n\}_{n=1}^{\infty}$  in which each function  $f_n$  is a mapping from  $X$  to  $\mathbb{R}$  *converges pointwise* to a function  $f$  on  $X$  if for every  $x \in X$ , the sequence of numbers  $\{f_n(x)\}_{n=0}^{\infty}$  converges to  $f(x)$ . On the other hand, a sequence of functions  $\{f_n\}_{n=1}^{\infty}$  *converges uniformly* to a function  $f$  on  $X$  if the following sequence  $\{M_n \mid M_n = \sup_{x \in X} |f_n(x) - f(x)|\}_{n=0}^{\infty}$  converges to 0.

Let us consider a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  where  $\Omega$  is a sample space,  $\mathcal{F}$  is a  $\sigma$ -algebra, and  $\mathbb{P}$  is a probability measure. A subset  $A$  of  $\mathcal{F}$  is called an event. The complement of an event  $A$  is denoted as  $A^c$ . Given a sequence of events  $\{A_n\}_{n=0}^{\infty}$ , we define  $\limsup_{n \rightarrow \infty} A_n$  as  $\bigcap_{n=0}^{\infty} \bigcup_{k=n}^{\infty} A_k$ , i.e. the event that  $A_n$  occurs infinitely often. In addition, the event  $\liminf_{n \rightarrow \infty} A_n$  is defined as  $\bigcup_{n=0}^{\infty} \bigcap_{k=n}^{\infty} A_k$ . A random variable is a measurable function mapping from  $\Omega$  to  $\mathbb{R}$ . The expected value of a random variable  $Y$  is defined as  $\mathbb{E}[Y] = \int_{\Omega} Y d\mathbb{P}$ . A sequence of random variables  $\{Y_n\}_{n=0}^{\infty}$  *converges surely* to a random variable  $Y$  if  $\lim_{n \rightarrow \infty} Y_n(\omega) = Y(\omega)$  for all  $\omega \in \Omega$ . A sequence of random variables  $\{Y_n\}_{n=0}^{\infty}$  *converges almost surely* or *with probability one* (w.p.1) to a random variable  $Y$  if  $\mathbb{P}(\omega \in \Omega \mid \lim_{n \rightarrow \infty} Y_n(\omega) = Y(\omega)) = 1$ . Almost sure convergence of  $\{Y_n\}_{n=0}^{\infty}$  to  $Y$  is denoted as  $Y_n \xrightarrow{a.s.} Y$ . We say that a sequence of random variables  $\{Y_n\}_{n=0}^{\infty}$  *converges in distribution* to a random variable  $Y$  if  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$  for every  $x \in \mathbb{R}$  at which  $F$  is continuous where  $\{F_n\}_{n=0}^{\infty}$  and  $F$  are the associated CDFs of  $\{Y_n\}_{n=0}^{\infty}$  and  $Y$  respectively. We

denote this convergence as  $Y_n \xrightarrow{d} Y$ . Convergence in distribution is also called weak convergence. If  $Y_n \xrightarrow{d} Y$ , then  $\lim_{n \rightarrow \infty} \mathbb{E}[f(Y_n)] = \mathbb{E}[f(Y)]$  for all bounded continuous functions  $f$ . As a corollary, when  $\{Y_n\}_{n=0}^\infty$  converges in distribution to 0, and  $Y_n$  is bounded for all  $n$ , we have  $\lim_{n \rightarrow \infty} \mathbb{E}[Y_n] = 0$  and  $\lim_{n \rightarrow \infty} \mathbb{E}[Y_n^2] = 0$ , which together imply  $\lim_{n \rightarrow \infty} \text{Var}(Y_n) = 0$ . We say that a sequence of random variables  $\{Y_n\}_{n=0}^\infty$  *converges in probability* to a random variable  $Y$ , denoted as  $Y_n \xrightarrow{p} Y$ , if for every  $\epsilon > 0$ , we have  $\lim_{n \rightarrow \infty} \mathbb{P}(|Y_n - Y| \geq \epsilon) = 0$ . For every continuous function  $f(\cdot)$ , if  $Y_n \xrightarrow{p} Y$ , then we also have  $f(Y_n) \xrightarrow{p} f(Y)$ . If  $Y_n \xrightarrow{p} Y$  and  $Z_n \xrightarrow{p} Z$ , then  $(Y_n, Z_n) \xrightarrow{p} (Y, Z)$ . If  $|Z_n - Y_n| \xrightarrow{p} 0$  and  $Y_n \xrightarrow{d} Y$ , we have  $Z_n \xrightarrow{d} Y$ . Finally, we say that a sequence of random variables  $\{Y_n\}_{n=0}^\infty$  *converges in  $r^{\text{th}}$  mean* to a random variable  $Y$ , denoted as  $Y_n \xrightarrow{r} Y$ , if  $\mathbb{E}[|X_n|^r] < \infty$  for all  $n$ , and  $\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - Y|^r] = 0$ . We have the following implications: (i) almost sure convergence or  $r^{\text{th}}$  mean convergence ( $r \geq 1$ ) implies convergence in probability, and (ii) convergence in probability implies convergence in distribution. The above results still hold for random vectors in higher dimensional spaces.

Let  $f(n)$  and  $g(n)$  be two functions with domain and range  $\mathbb{N}$  or  $\mathbb{R}$ . The function  $f(n)$  is called  $O(g(n))$  if there exists two constants  $M$  and  $n_0$  such that  $f(n) \leq Mg(n)$  for all  $n \geq n_0$ . The function  $f(n)$  is called  $\Omega(g(n))$  if  $g(n)$  is  $O(f(n))$ . Finally, the function  $f(n)$  is called  $\Theta(g(n))$  if  $f(n)$  is both  $O(g(n))$  and  $\Omega(g(n))$ .

## B Proof of Lemma 4

For each  $n \in \mathbb{N}$ , divide the state space  $S$  into grid cells with side length  $1/2\gamma_r(\log |S_n|/|S_n|)^{1/d_x}$  as follows. Let  $\mathbb{Z}$  denote the set of integers. Define the grid cell  $i \in \mathbb{Z}^{d_x}$  as

$$W_n(i) := i \left( \frac{\gamma_r \log |S_n|}{2 |S_n|} \right)^{1/d_x} + \left[ -\frac{1}{4} \gamma_r \left( \frac{\log |S_n|}{|S_n|} \right)^{1/d_x}, \frac{1}{4} \gamma_r \left( \frac{\log |S_n|}{|S_n|} \right)^{1/d_x} \right]^{d_x},$$

where  $[-a, a]^{d_x}$  denotes the  $d_x$ -dimensional cube with side length  $2a$  centered at the origin. Hence, the expression above translates the  $d_x$ -dimensional cube with side length  $(1/2) \gamma_r (\log |S_n|/|S_n|)^{1/d_x}$  to the point with coordinates  $i \frac{\gamma_r}{2} (\log |S_n|/|S_n|)^{1/d_x}$ .

Let  $Q_n$  denote the indices of set of all cells that lie completely inside the state space  $S$ , i.e.,  $Q_n = \{i \in \mathbb{Z}^{d_x} : W_n(i) \subseteq S\}$ . Clearly,  $Q_n$  is finite since  $S$  is bounded. Let  $\partial Q_n$  denote the set of all grid cells that intersect the boundary of  $S$ , i.e.,  $\partial Q_n = \{i \in \mathbb{Z}^{d_x} : W_n(i) \cap \partial S \neq \emptyset\}$ . We claim for all large  $n$ , all grid cells in  $Q_n$  contain one vertex of  $S_n$ , and all grid cells in  $\partial Q_n$  contain one vertex from  $\partial S_n$ . First, let us show that each cell in  $Q_n$  contains at least one vertex. Given an event  $A$ , let  $A^c$  denote its complement. Let  $A_{n,k}$  denote the event that the cell  $W_n(k)$ , where  $k \in Q_n$  contains a vertex from  $S_n$ , and let  $A_n$  denote the event that all grid cells in  $Q_n$  contain a vertex in  $S_n$ . Then, for all  $k \in Q_n$ ,

$$\mathbb{P}(A_{n,k}^c) = \left( 1 - \frac{(\gamma_r/2)^{d_x} \log |S_n|}{m(S) |S_n|} \right)^{|S_n|} \leq \exp \left( -((\gamma_r/2)^{d_x}/m(S)) \log |S_n| \right) = |S_n|^{-(\gamma_r/2)^{d_x}/m(S)},$$

where  $m(S)$  denotes Lebesgue measure assigned to  $S$ . Then,

$$\mathbb{P}(A_n^c) = \mathbb{P} \left( \left( \bigcap_{k \in Q_n} A_{n,k} \right)^c \right) = \mathbb{P} \left( \bigcup_{k \in Q_n} A_{n,k}^c \right) \leq \sum_{k \in Q_n} \mathbb{P}(A_{n,k}^c) = |Q_n| |S_n|^{-(\gamma_r/2)^{d_x}/m(S)},$$

where the first inequality follows from the union bound and  $|Q_n|$  denotes the cardinality of the set  $Q_n$ . By calculating the maximum number of cubes that can fit into  $S$ , we can bound  $|Q_n|$ :

$$|Q_n| \leq \frac{m(S)}{(\gamma_r/2)^{d_x} \frac{\log |S_n|}{|S_n|}} = \frac{m(S)}{(\gamma_r/2)^{d_x}} \frac{|S_n|}{\log |S_n|}.$$

Note that by construction, we have  $|S_n| = \Theta(n)$ . Thus,

$$\begin{aligned}\mathbb{P}(A_n^c) &\leq \frac{m(S)}{(\gamma_r/2)^{d_x}} \frac{|S_n|}{\log |S_n|} |S_n|^{-(\gamma_r/2)^{d_x}/m(S)} = \frac{m(S)}{(\gamma_r/2)^{d_x}} \frac{1}{\log |S_n|} |S_n|^{1-(\gamma_r/2)^{d_x}/m(S)} \\ &\leq \frac{m(S)}{(\gamma_r/2)^{d_x}} |S_n|^{1-(\gamma_r/2)^{d_x}/m(S)},\end{aligned}$$

which is summable for all  $\gamma_r > 2(2m(S))^{1/d_x}$ . Hence, by the Borel-Cantelli lemma, the probability that  $A_n^c$  occurs infinitely often is zero, which implies that the probability that  $A_n$  occurs for all large  $n$  is one, i.e.,  $\mathbb{P}(\liminf_{n \rightarrow \infty} A_n) = 1$ .

Similarly, each grid cell in  $\partial Q_n$  can be shown to contain at least one vertex from  $\partial S_n$  for all large  $n$ , with probability one. This implies each grid cell in both sets  $Q_n$  and  $\partial Q_n$  contain one vertex of  $S_n$  and  $\partial S_n$ , respectively, for all large  $n$ , with probability one. Hence the following event happens with probability one:

$$\zeta_n = \max_{z \in S_n} \min_{z' \in S_n} \|z' - z\|_2 = O((\log |S_n|/|S_n|)^{1/d_x}).$$

□

## C Proof of Lemma 5

We show that each state that is added to the approximating MDPs is updated infinitely often. That is, for any  $z \in S_n$ , the set of all iterations in which the procedure **Update** is applied on  $z$  is unbounded. Indeed, let us denote  $\zeta_n(z) = \min_{z' \in S_n} \|z' - z\|_2$ . From Lemma 4,  $\lim_{n \rightarrow \infty} \zeta_n(z) = 0$  happens almost surely. Therefore, with probability one, there are infinitely many  $n$  such that  $\zeta_n(z) < \zeta_{n-1}(z)$ . In other words, with probability one, we can find infinitely many  $z_{new}$  at Line 13 of Algorithm 1 such that  $z$  is updated. For those  $n$ , the holding time at  $z$  is recomputed as  $\Delta t_n(z) = \gamma_t \left( \frac{\log |S_n|}{|S_n|} \right)^{\theta \varsigma \rho / d_x}$  at Line 1 of Algorithm 2. Thus, the following event happens with probability one:

$$\lim_{n \rightarrow \infty} \Delta t_n(z) = 0,$$

which satisfies the first condition of local consistency in Eq. 3.

The other conditions of local consistency in Eqs. 4-6 are satisfied immediately by the way that the transition probabilities are computed (see the description of the **ComputeTranProb** procedure given in Section 4.1). Hence, the MDP sequence  $\{\mathcal{M}_n\}_{n=0}^\infty$  and holding times  $\{\Delta t_n\}_{n=0}^\infty$  are locally consistent for large  $n$  with probability one. □

## D Proof of Theorem 7

To highlight the idea of the entire proof, we first prove the convergence under synchronous value iterations before presenting the convergence under asynchronous value iterations. As we will see, the shrinking rate of holding times plays a crucial role in the convergence proof. The outline of the proof is as follows.

- S1: Convergence under synchronous value iterations: In Algorithm 1, we take  $L_n \geq 1$  and  $K_n = |S_n| - 1$ . In other words, in each iteration, we perform synchronous value iterations. Moreover, we assume that we are able to solve the Bellman equation (Eq. 9) exactly. We show that  $J_n$  converges uniformly to  $J^*$  almost surely in this setting.

S2: Convergence under asynchronous value iterations: When  $K_n = \Theta(|S_n|^\theta) < |S_n|$ , we only update a subset of  $S_n$  in each of  $L_n$  passes. We show that  $J_n$  still converges uniformly to  $J^*$  almost surely in this new setting.

In the following discussion and next sections, we need to compare functions on different domains  $S_n$ . To ease the discussion and simplify the notation, we adopt the following interpolation convention. Given  $X \subset Y$  and  $J : X \rightarrow \mathbb{R}$ , we interpolate  $J$  to  $\bar{J}$  on the entire domain  $Y$  via nearest neighbor value:

$$\forall y \in Y : \quad \bar{J}(y) = J(z) \text{ where } z = \operatorname{argmin}_{z' \in X} \|z' - y\|.$$

To compare  $J : X \rightarrow \mathbb{R}$  and  $J' : Y \rightarrow \mathbb{R}$  where  $X, Y \subset S$ , we define the sup-norm:

$$\|J - J'\|_\infty = \|\bar{J} - \bar{J}'\|_\infty,$$

where  $\bar{J}$  and  $\bar{J}'$  are interpolations of  $J$  and  $J'$  from the domains  $X$  and  $Y$  to the entire domain  $S$  respectively. In particular, given  $J_n : S_n \rightarrow \mathbb{R}$ , and  $J : S \rightarrow \mathbb{R}$ , then  $\|J_n - J\|_{S_n} \leq \|J_n - J\|_\infty$ . Thus, if  $\|J_n - J\|_\infty$  approaches 0 when  $n$  approaches  $\infty$ , so does  $\|J_n - J\|_{S_n}$ . Hence, we will work with the (new) sup-norm  $\|\cdot\|_\infty$  instead of  $\|\cdot\|_{S_n}$  in the proofs of Theorems 7-8. The triangle inequality also holds for any functions  $J, J', J''$  defined on subsets of  $S$  with respect to the above sup-norm:

$$\|J - J'\|_\infty \leq \|J - J''\|_\infty + \|J'' - J'\|_\infty.$$

Let  $B(X)$  denote a set of all real-valued bounded functions over a domain  $X$ . For  $S_n \subset S_{n'}$  when  $n < n'$ , a function  $J$  in  $B(S_n)$  also belongs to  $B(S_{n'})$ , meaning that we can interpolate  $J$  on  $S_n$  to a function  $J'$  on  $S_{n'}$ . In particular, we say that  $J$  in  $B(S_n)$  also belongs to  $B(S)$ .

Lastly, due to random sampling,  $S_n$  is a random set, and therefore functions  $J_n$  and  $J_n^*$  defined on  $S_n$  are random variables. In the following discussion, inequalities hold surely without further explanation when it is clear from the context, and inequalities hold almost surely if they are followed by “w.p.1”.

## S1: Convergence under synchronous value iterations

In this step, we first set  $L_n \geq 1$  and  $K_n = |S_n| - 1$  in Algorithm 1. Thus, for all  $z \in S_n$ , the holding time  $\Delta t_n(z)$  equals  $\gamma_t \left( \frac{\log |S_n|}{|S_n|} \right)^{\theta \zeta \rho / d_x}$  and is denoted as  $\Delta t_n$ . We consider the MDP  $\mathcal{M}_n = (S_n, U, P_n, G_n, H_n)$  at  $n^{\text{th}}$  iteration and define the following operator  $T_n : B(S_n) \rightarrow B(S_n)$  that transforms every  $J \in B(S_n)$  after a Bellman update as:

$$T_n J(z) = \min_{v \in U} \{G_n(z, v) + \alpha^{\Delta t_n} \mathbb{E}_{P_n} [J(y)|z, v]\}, \quad \forall z \in S_n, \quad (10)$$

assuming that we can solve the minimization on the RHS of Eq. 10 exactly. For each  $k \geq 2$ , operators  $T_n^k$  are defined recursively as  $T_n^k = T_n T_n^{k-1}$  and  $T_n^1 = T_n$ . When we apply  $T_n$  on  $J \in B(S_k)$  where  $k < n$ ,  $J$  is interpolated to  $S_n$  before applying  $T_n$ . Thus, in Algorithms 1-2, we implement the next update

$$J_n = T_n^{L_n} J_{n-1}.$$

From [26], we have the following results:  $J_n^* = T_n J_n^*$ , and  $T_n$  is a contraction mapping. For any  $J$  and  $J'$  in  $B(S_n)$ , the following inequality happens surely:

$$\|T_n J - T_n J'\|_\infty \leq \alpha^{\Delta t_n} \|J - J'\|_\infty.$$

Combining the above results:

$$\begin{aligned} \|J_n^* - J_n\|_\infty &= \|T_n^{L_n} J_n^* - T_n^{L_n} J_{n-1}\|_\infty \leq \alpha^{L_n \Delta t_n} \|J_n^* - J_{n-1}\|_\infty \\ &\leq \alpha^{\Delta t_n} (\|J_n^* - J_{n-1}^*\|_\infty + \|J_{n-1}^* - J_{n-1}\|_\infty), \end{aligned}$$

where the second inequality follows from the triangle inequality, and  $L_n \geq 1, \alpha \in (0, 1)$ .

Thus, by iterating over  $n$ , for any  $N \geq 1$  and  $n > N$ , we have:

$$\|J_n^* - J_n\|_\infty \leq A_n + \alpha^{\Delta t_n + \Delta t_{n-1} \dots + \Delta t_{N+1}} \|J_N^* - J_N\|_\infty, \quad (11)$$

where  $A_n$  are defined recursively:

$$A_n = \alpha^{\Delta t_n} (\|J_n^* - J_{n-1}^*\|_\infty + A_{n-1}), \quad \forall n > N + 1, \quad (12)$$

$$A_{N+1} = \alpha^{\Delta t_{N+1}} \|J_{N+1}^* - J_N^*\|_\infty. \quad (13)$$

Note that for any  $N \geq 1$ :

$$\lim_{n \rightarrow \infty} \Delta t_n + \Delta t_{n-1} \dots + \Delta t_{N+1} = \infty,$$

due to the choice of holding times  $\Delta t_n$  in the procedure **ComputeHoldingTime**. Therefore,

$$\lim_{n \rightarrow \infty} \alpha^{\Delta t_n + \dots + \Delta t_{N+1}} \|J_N^* - J_N\|_\infty = 0.$$

By Theorem 6, the following event happens with probability 1 (w.p.1):

$$\lim_{n \rightarrow \infty} \|J_n^* - J^*\|_\infty = 0,$$

hence,

$$\lim_{n \rightarrow \infty} \|J_n^* - J_{n-1}^*\|_\infty = 0 \text{ w.p.1.}$$

Thus, for any fixed  $\epsilon > 0$ , we can choose  $N$  large enough such that:

$$\|J_n^* - J_{n-1}^*\|_\infty^{1-\varsigma} < \epsilon \text{ w.p.1 for all } n > N, \text{ and} \quad (14)$$

$$\alpha^{\Delta t_n + \dots + \Delta t_{N+1}} \|J_N^* - J_N\|_\infty < \epsilon \text{ surely,} \quad (15)$$

where  $\varsigma \in (0, 1)$  is the constant defined in the procedure **ComputeHoldingTime**.

Now, for all  $n > N$ , we rearrange Eqs.12-13 to have

$$A_n \leq \epsilon B_n \text{ w.p.1,}$$

where

$$\begin{aligned} B_n &= \alpha^{\Delta t_n} (\|J_n^* - J_{n-1}^*\|_\infty^\varsigma + B_{n-1}), \quad \forall n > N + 1, \\ B_{N+1} &= \alpha^{\Delta t_{N+1}} \|J_{N+1}^* - J_N^*\|_\infty^\varsigma. \end{aligned}$$

We can see that for  $n > N + 1$ :

$$B_n = \alpha^{\Delta t_n} (\|J_n^* - J_{n-1}^*\|_\infty^\varsigma + B_{n-1}) < \epsilon^{\varsigma/(1-\varsigma)} + B_{n-1} \text{ w.p.1,} \quad (16)$$

$$B_{N+1} = \alpha^{\Delta t_{N+1}} \|J_{N+1}^* - J_N^*\|_\infty^\varsigma < \epsilon^{\varsigma/(1-\varsigma)} \text{ w.p.1.} \quad (17)$$



We now prove that almost surely,  $B_n$  is bounded for all  $n \geq N$ . Indeed, we derive the conditions so that  $B_{n-1} < B_n$  or  $B_{n-1} \geq B_n$  as follows:

$$\begin{aligned}
& B_{n-1} < B_n \\
& \Leftrightarrow B_{n-1} < \alpha^{\Delta t_n} (\|J_n^* - J_{n-1}^*\|_\infty^\varsigma + B_{n-1}) \\
& \Leftrightarrow B_{n-1} < \frac{\alpha^{\Delta t_n} \|J_n^* - J_{n-1}^*\|_\infty^\varsigma}{1 - \alpha^{\Delta t_n}} \\
& \Rightarrow B_{n-1} < \mathcal{K} \frac{\alpha^{\gamma_t \left(\frac{\log |S_n|}{|S_n|}\right)^{\theta \varsigma \rho / d_x}} \left(\frac{\log |S_n|}{|S_n|}\right)^{\varsigma \rho / d_x}}{1 - \alpha^{\gamma_t \left(\frac{\log |S_n|}{|S_n|}\right)^{\theta \varsigma \rho / d_x}}} \text{ w.p.1.}
\end{aligned}$$

The last inequality is due to Theorem 6 and  $|S_n| = \Theta(n)$ ,  $|S_{n-1}| = \Theta(n-1)$ :

$$\|J_n^* - J_{n-1}^*\|_\infty = O((\log |S_{n-1}| / |S_{n-1}|)^{\rho / d_x}) < \mathcal{K} \left(\frac{\log |S_n|}{|S_n|}\right)^{\rho / d_x} \text{ w.p.1,}$$

for large  $n$  where  $\mathcal{K}$  is some finite constant. Let  $\beta = \alpha^{\gamma_t} \in (0, 1)$ . For large  $n$ ,  $\frac{\log |S_n|}{|S_n|}$  are in  $(0, 1)$  and  $\theta \in (0, 1]$ . Let us define

$$x_n = \left(\frac{\log |S_n|}{|S_n|}\right)^{\theta \varsigma \rho / d_x}, \text{ and } y_n = \left(\frac{\log |S_n|}{|S_n|}\right)^{\varsigma \rho / d_x}.$$

Then,  $x_n \geq y_n > 0$ . The above condition is simplified to

$$B_{n-1} < \mathcal{K} \frac{\beta^{x_n} y_n}{1 - \beta^{x_n}} \leq \mathcal{K} \frac{\beta^{x_n} x_n}{1 - \beta^{x_n}}, \text{ w.p.1.}$$

Consider the function  $r : [0, \infty) \rightarrow \mathbb{R}$  such that  $r(x) = \frac{\beta^x x}{1 - \beta^x}$ , we can verify that  $r(x)$  is non-increasing and is bounded by  $r(0) = -1 / \log(\beta)$ . Therefore:

$$B_{n-1} < B_n \Rightarrow B_{n-1} < -\frac{\mathcal{K}}{\log(\beta)} = -\frac{\mathcal{K}}{\gamma_t \log(\alpha)} \text{ w.p.1.} \quad (18)$$

Or conversely,

$$B_{n-1} \geq -\frac{\mathcal{K}}{\gamma_t \log(\alpha)} \text{ w.p.1} \Rightarrow B_{n-1} \geq B_n \text{ w.p.1.} \quad (19)$$

The above discussion characterizes the random sequence  $B_n$ . In particular, Fig. 5 shows a possible realization of the random sequence  $B_n$  for  $n > N$ . As shown visually in this plot,  $B_{N+1}$  is less than  $\epsilon^{\varsigma/(1-\varsigma)}$  w.p.1 and thus is less than  $\epsilon^{\varsigma/(1-\varsigma)} - \frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1. For  $n > N + 1$ , assume that we have already shown that  $B_{n-1}$  is bounded from above by  $\epsilon^{\varsigma/(1-\varsigma)} - \frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1. When  $B_{n-1} \geq -\frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1, the sequence is non-increasing w.p.1. Conversely, when the sequence is increasing, i.e.  $B_{n-1} < B_n$ , we assert that  $B_{n-1} < -\frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1 due to Eq. 18, and the increment is less than  $\epsilon^{\varsigma/(1-\varsigma)}$  due to Eq. 16. In both cases, we conclude that  $B_n$  is also bounded by  $\epsilon^{\varsigma/(1-\varsigma)} - \frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1. Hence, from Eqs. 16-19, we infer that  $B_n$  is bounded w.p.1 for all  $n > N$ :

$$B_n < \epsilon^{\varsigma/(1-\varsigma)} - \frac{\mathcal{K}}{\gamma_t \log(\alpha)} \text{ w.p.1.}$$

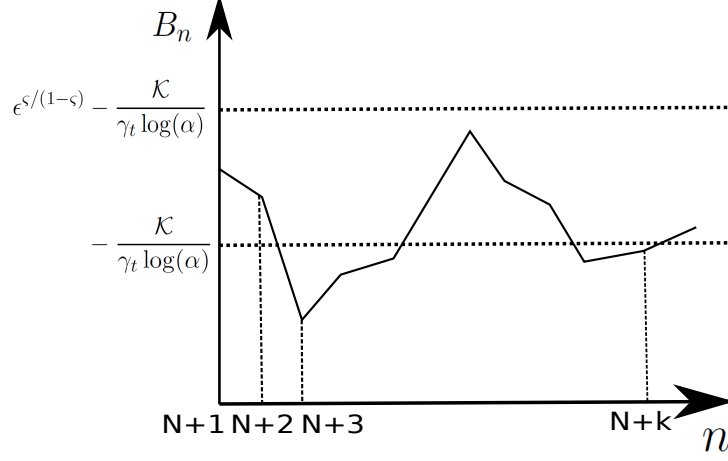


Figure 5: A realization of the random sequence  $B_n$ . We have  $B_{N+1}$  less than  $\epsilon^{\varsigma/(1-\varsigma)}$  w.p.1. For  $n$  larger than  $N + 1$ , when  $B_{n-1} \geq -\frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1, the sequence is non-increasing w.p.1, i.e.  $B_{n-1} \geq B_n$  w.p.1. Conversely, when the sequence is increasing, i.e.  $B_{n-1} < B_n$ , we have  $B_{n-1} < -\frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1, and the increment is less than  $\epsilon^{\varsigma/(1-\varsigma)}$ . Hence, the random sequence  $B_n$  is bounded by  $\epsilon^{\varsigma/(1-\varsigma)} - \frac{\mathcal{K}}{\gamma_t \log(\alpha)}$  w.p.1.

Thus, for all  $n > N$ :

$$A_n \leq \epsilon B_n < \epsilon \left( \epsilon^{\varsigma/(1-\varsigma)} - \frac{\mathcal{K}}{\gamma_t \log(\alpha)} \right) \text{ w.p.1.} \quad (20)$$

Combining Eqs. 11,15, and 20, we conclude that for any  $\epsilon > 0$ , there exists  $N \geq 1$  such that for all  $n > N$ , we have

$$\|J_n^* - J_n\|_\infty < \epsilon \left( \epsilon^{\varsigma/(1-\varsigma)} - \frac{\mathcal{K}}{\gamma_t \log(\alpha)} + 1 \right) \text{ w.p.1.}$$

Therefore,

$$\lim_{n \rightarrow \infty} \|J_n^* - J_n\|_\infty = 0 \text{ w.p.1.}$$

Combining with

$$\lim_{n \rightarrow \infty} \|J_n^* - J^*\|_\infty = 0 \text{ w.p.1,}$$

we obtain

$$\lim_{n \rightarrow \infty} \|J_n - J^*\|_\infty = 0 \text{ w.p.1.}$$

In the above analysis, the shrinking rate  $\left( \frac{\log |S_n|}{|S_n|} \right)^{\theta \varsigma \rho / d_x}$  of holding times plays an important role to construct an upper bound of the sequence  $B_n$ . This rate must be slower than the convergence rate  $\left( \frac{\log |S_n|}{|S_n|} \right)^{\rho / d_x}$  of  $J_n^*$  to  $J^*$  so that the function  $r(x)$  is bounded, enabling the convergence of cost value functions  $J_n$  to the optimal cost-to-go  $J^*$ . Remarkably, we have accomplished this convergence by carefully selecting the range  $(0, 1)$  of the parameter  $\varsigma$ . The role of the parameter  $\theta$  in this convergence will be clear in Step S2. Lastly, we note that if we are able to obtain a faster convergence rate of  $J_n^*$  to  $J^*$ , we can have faster shrinking rate for holding times.

## S2: Convergence under asynchronous value iterations

When  $1 \leq L_n$  and  $K_n = \Theta(|S_n|^\theta) < |S_n|$ , we first claim the following result:

**Lemma 10** Consider any increasing sequence  $\{n_k\}_{k=0}^\infty$  as a subset of  $\mathbb{N}$  such that  $n_0 = 0$  and  $k \leq |S_{n_k}| \leq k^{1/\theta}$ . For  $J \in B(S)$ , we define:

$$A(\{n_j\}_{j=0}^k) = \alpha^{\Delta t_{n_k} + \Delta t_{n_{k-1}} + \dots + \Delta t_{n_1}} \|J_{n_1}^* - J\|_\infty + \alpha^{\Delta t_{n_k} + \Delta t_{n_{k-1}} + \dots + \Delta t_{n_2}} \|J_{n_2}^* - J_{n_1}^*\|_\infty \\ + \dots + \alpha^{\Delta t_{n_k}} \|J_{n_k}^* - J_{n_{k-1}}^*\|_\infty.$$

The following event happens with probability one:

$$\lim_{k \rightarrow \infty} A(\{n_j\}_{j=0}^k) = 0.$$

**Proof** We rewrite  $A(\{n_j\}_{j=0}^k) = A_{n_k}$  where  $A_{n_k}$  are defined recursively:

$$A_{n_k} = \alpha^{\Delta t_{n_k}} (\|J_{n_k}^* - J_{n_{k-1}}^*\|_\infty + A_{n_{k-1}}), \quad \forall k > K, \quad (21)$$

$$A_{n_K} = A(\{n_j\}_{j=0}^K), \quad \forall K \geq 1. \quad (22)$$

We note that

$$\Delta t_{n_k} + \Delta t_{n_{k-1}} + \dots + \Delta t_{n_K} \\ = \gamma_t \left( \frac{\log |S_{n_k}|}{|S_{n_k}|} \right)^{\theta \varsigma \rho / d_x} + \gamma_t \left( \frac{\log |S_{n_{k-1}}|}{|S_{n_{k-1}}|} \right)^{\theta \varsigma \rho / d_x} + \dots + \gamma_t \left( \frac{\log |S_{n_K}|}{|S_{n_K}|} \right)^{\theta \varsigma \rho / d_x} \\ \geq \gamma_t \left( \frac{1}{|S_{n_k}|} \right)^{\theta \varsigma \rho / d_x} + \gamma_t \left( \frac{1}{|S_{n_{k-1}}|} \right)^{\theta \varsigma \rho / d_x} + \dots + \gamma_t \left( \frac{1}{|S_{n_K}|} \right)^{\theta \varsigma \rho / d_x} \\ \geq \gamma_t \frac{1}{k^{\varsigma \rho / d_x}} + \gamma_t \frac{1}{(k-1)^{\varsigma \rho / d_x}} + \dots + \gamma_t \frac{1}{(K)^{\varsigma \rho / d_x}} \geq \gamma_t \left( \frac{1}{k} + \frac{1}{k-1} + \dots + \frac{1}{K} \right),$$

where the second inequality uses the given fact that  $|S_{n_k}| \leq k^{1/\theta}$ . Therefore, for any  $K \geq 1$ :

$$\lim_{k \rightarrow \infty} \alpha^{\Delta t_{n_k} + \Delta t_{n_{k-1}} + \dots + \Delta t_{n_K}} = 0.$$

We choose a constant  $\varrho > 1$  such that  $\varrho \varsigma < 1$ . For any fixed  $\epsilon > 0$ , we can choose  $K$  large enough such that:

$$\|J_{n_k}^* - J_{n_{k-1}}^*\|_\infty^{1-\varrho \varsigma} < \epsilon \text{ w.p.1 for all } k > K. \quad (23)$$

For all  $k > K$ , we can write

$$A_{n_k} \leq \epsilon B_{n_k} + \alpha^{\Delta t_{n_k} + \dots + \Delta t_{n_{K+1}}} A(\{n_j\}_{j=0}^K).$$

where

$$B_{n_k} = \alpha^{\Delta t_{n_k}} (\|J_{n_k}^* - J_{n_{k-1}}^*\|_\infty^{\varrho \varsigma} + B_{n_{k-1}}), \quad \forall k > K, \\ B_{n_K} = 0.$$

Furthermore, we can choose  $K'$  sufficiently large such that  $K' \geq K$  and for all  $k > K'$ :

$$\alpha^{\Delta t_{n_k} + \dots + \Delta t_{n_{K+1}}} A(\{n_j\}_{j=0}^K) \leq \epsilon.$$

We obtain:

$$A_{n_k} \leq \epsilon B_{n_k} + \epsilon, \quad \forall k > K' \geq K \geq 1.$$

We can also see that for  $k > K$ :

$$B_{n_k} = \alpha^{\Delta t_{n_k}} (\|J_{n_k}^* - J_{n_{k-1}}^*\|_\infty^{\varrho\varsigma} + B_{n_{k-1}}) < \epsilon^{\varrho\varsigma/(1-\varrho\varsigma)} + B_{n_{k-1}} \quad \text{w.p.1.} \quad (24)$$

Similar to Step S1, we characterize the random sequence  $B_{n_k}$  as follows:

$$\begin{aligned} B_{n_{k-1}} &< B_{n_k} \\ \Leftrightarrow B_{n_{k-1}} &< \frac{\alpha^{\Delta t_{n_k}} \|J_{n_k}^* - J_{n_{k-1}}^*\|_\infty^{\varrho\varsigma}}{1 - \alpha^{\Delta t_{n_k}}} \\ \Rightarrow B_{n_{k-1}} &< \mathcal{K} \frac{\alpha^{\gamma_t \left( \frac{\log |S_{n_k}|}{|S_{n_k}|} \right)^{\theta\varsigma\rho/d_x}} \left( \frac{\log |S_{n_{k-1}}|}{|S_{n_{k-1}}|} \right)^{\varrho\varsigma\rho/d_x}}{1 - \alpha^{\gamma_t \left( \frac{\log |S_{n_k}|}{|S_{n_k}|} \right)^{\theta\varsigma\rho/d_x}}} \quad \text{w.p.1.} \end{aligned}$$

Let  $\beta = \alpha^{\gamma_t} \in (0, 1)$ . We define:

$$x_k = \left( \frac{\log |S_{n_k}|}{|S_{n_k}|} \right)^{\theta\varsigma\rho/d_x}, \quad \text{and} \quad y_k = \left( \frac{\log |S_{n_{k-1}}|}{|S_{n_{k-1}}|} \right)^{\varrho\varsigma\rho/d_x}.$$

We note that  $\frac{\log x}{x}$  is a decreasing function for positive  $x$ . Since  $|S_{n_{k-1}}| \geq k-1$  and  $|S_{n_k}| \leq k^{1/\theta}$ , we have the following inequalities:

$$x_k \geq \left( \frac{(\frac{\log k}{\theta})^\theta}{k} \right)^{\varsigma\rho/d_x}, \quad y_k \leq \left( \frac{(\log(k-1))^\varrho}{(k-1)^\varrho} \right)^{\varsigma\rho/d_x}.$$

Since  $\theta \in (0, 1]$  and  $\varrho > 1$ , we can find a finite constant  $\mathcal{K}_1$  such that  $y_k < \mathcal{K}_1 x_k$  for large  $k$ . Thus, the above condition leads to

$$B_{n_{k-1}} < \mathcal{K} \frac{\beta^{x_k} y_k}{1 - \beta^{x_k}} < \mathcal{K} \mathcal{K}_1 \frac{\beta^{x_k} x_k}{1 - \beta^{x_k}}, \quad \text{w.p.1.}$$

Therefore:

$$B_{n_{k-1}} < B_{n_k} \quad \Rightarrow \quad B_{n_{k-1}} < -\frac{\mathcal{K} \mathcal{K}_1}{\log(\beta)} = -\frac{\mathcal{K} \mathcal{K}_1}{\gamma_t \log(\alpha)} \quad \text{w.p.1.}$$

Or conversely,

$$B_{n_{k-1}} \geq -\frac{\mathcal{K} \mathcal{K}_1}{\gamma_t \log(\alpha)} \quad \text{w.p.1} \quad \Rightarrow \quad B_{n-1} \geq B_n \quad \text{w.p.1.}$$

Arguing similarly to Step S1, we infer that for all  $k > K' \geq K \geq 1$ :

$$B_{n_k} < \epsilon^{\varrho\varsigma/(1-\varrho\varsigma)} - \frac{\mathcal{K} \mathcal{K}_1}{\gamma_t \log(\alpha)} \quad \text{w.p.1.}$$

Thus, for any  $\epsilon > 0$ , we can find  $K' \geq 1$  such that for all  $k > K'$ :

$$A_{n_k} \leq \epsilon B_{n_k} + \epsilon < \epsilon \left( \epsilon^{\varrho\varsigma/(1-\varrho\varsigma)} - \frac{\mathcal{K} \mathcal{K}_1}{\gamma_t \log(\alpha)} + 1 \right) \quad \text{w.p.1.}$$

We conclude that

$$\lim_{k \rightarrow \infty} A(\{n_j\}_{j=0}^k) = 0. \quad \text{w.p.1.}$$

□

Returning to the main proof, we use the tilde notation to indicate asynchronous operations to differentiate with our synchronous operations in Step S1. We will also assume that  $L_n = 1$  for all  $n$  to simplify the following notations. The proof for general  $L_n \geq 1$  is exactly the same. We define the following (asynchronous) mappings  $\tilde{T}_n : B(S_n) \rightarrow B(S_n)$  as the restricted mappings of  $T_n$  on  $D_n$ , a non-empty random subset of  $S_n$ , such that for all  $J \in B(S_n)$ :

$$\tilde{T}_n J(z) = \min_{v \in U} \left\{ G_n(z, v) + \alpha^{\Delta t_n} \mathbb{E}_{P_n} [J(y)|z, v] \right\}, \quad \forall z \in D_n \subset S_n, \quad (25)$$

$$\tilde{T}_n J(z) = J(z), \quad \forall z \in S_n \setminus D_n. \quad (26)$$

We require that

$$\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} D_k = S. \quad (27)$$

In other words, every state in  $S$  are sampled infinitely often. We can see that in Algorithm 1, if the set  $Z_{\text{update}}$  is assigned to  $D_n$  in every iteration (Line 13), the sequence  $\{D_n\}_{n=1}^{\infty}$  has the above property, and  $|D_n| = \Theta(|S_n|^\theta) < |S_n|$ .

Starting from any  $\tilde{J}_0 \in B(S_0)$ , we perform the following asynchronous iteration

$$\tilde{J}_{n+1} = \tilde{T}_{n+1} \tilde{J}_n, \quad \forall n \geq 0. \quad (28)$$

Consider the following sequence  $\{m_k\}_{k=0}^{\infty}$  such that  $m_0 = 0$  and for all  $k \geq 0$ , from  $m_k$  to  $m_{k+1} - 1$ , all states in  $S_{m_{k+1}-1}$  are chosen to be updated at least once, and a subset of states in  $S_{m_{k+1}-1}$  is chosen to be updated exactly once. We observe that as the size of  $S_n$  increases linearly with  $n$ , if we schedule states in  $D_n \subset S_n$  to be updated in a round-robin manner, we have  $k \leq S_{m_k} \leq k^{1/\theta}$ . When  $D_n$  is chosen as shown in Algorithm 1, with high probability,  $k \leq S_{m_k} \leq k^{1/\theta}$ . However, we will assume that the event  $k \leq S_{m_k} \leq k^{1/\theta}$  happens surely because we can always schedule a fraction of  $D_n$  to be updated in a round-robin manner.

We define  $W_n$  as the set of increasing sub-sequences of the sequence  $\{0, 1, \dots, n\}$  such that each sub-sequence contains  $\{m_j\}_{j=0}^k$  where  $m_k \leq n < m_{k+1}$ :

$$W_n = \left\{ \{i_j\}_{j=0}^T \mid \{m_j\}_{j=0}^k \subset \{i_j\}_{j=0}^T \subset \{0, 1, \dots, n\} \wedge T \geq 2 \wedge m_k \leq n < m_{k+1} \right\}.$$

Clearly, if  $\{i_j\}_{j=0}^T \in W_n$ , we have  $i_0 = 0$ . For each  $\{i_j\}_{j=0}^T \in W_n$ , we define

$$\begin{aligned} A(\{i_j\}_{j=0}^T) &= \alpha^{\Delta t_{i_T} + \Delta t_{i_{T-1}} + \dots + \Delta t_{i_1}} \|J_{i_1}^* - \tilde{J}_0\|_{\infty} + \alpha^{\Delta t_{i_T} + \Delta t_{i_{T-1}} + \dots + \Delta t_{i_2}} \|J_{i_2}^* - J_{i_1}^*\|_{\infty} \\ &\quad + \dots + \alpha^{\Delta t_{i_T}} \|J_{i_T}^* - J_{i_{T-1}}^*\|_{\infty}. \end{aligned}$$

We will prove by induction that

$$\forall z \in D_n \Rightarrow |\tilde{J}_n(z) - J_n^*(z)| \leq \max_{\{i_j\}_{j=0}^T \in W_n} A(\{i_j\}_{j=0}^T). \quad (29)$$

When  $n = 1$ , the only sub-sequence is  $\{i_j\}_{j=0}^T = \{0, 1\} \in W_1$ . It is clear that for  $z \in D_1$ , due to the contraction property of  $T_1$ :

$$|J_1^*(z) - \tilde{J}_1(z)| \leq \max_{\{i_j\}_{j=0}^T \in W_1} A(\{i_j\}_{j=0}^T) = \alpha^{\Delta t_1} \|J_1^* - \tilde{J}_0\|_{\infty}.$$

Assuming that Eq. 29 holds upto  $n = m_k$ , we need to prove that the equation also holds for those  $n \in (m_k, m_{k+1})$  and  $n = m_{k+1}$ . Indeed, let us assume that Eq. 29 holds for some  $n \in [m_k, m_{k+1}-1]$ .

Denote  $n_z \leq n$  as the index of the most recent update of  $z$ . For  $z \in D_n$ , we compute new values for  $z$  in  $\tilde{J}_{n+1}$ , and by the contraction property of  $T_{n+1}$ , it follows that

$$\begin{aligned}
|\tilde{J}_{n+1}(z) - J_{n+1}^*(z)| &\leq \alpha^{\Delta t_{n+1}} \|J_{n+1}^* - \tilde{J}_n\|_\infty \\
&= \alpha^{\Delta t_{n+1}} \max_{z \in S_{n+1}} |J_{n+1}^*(z) - \tilde{J}_n(z)| \\
&= \alpha^{\Delta t_{n+1}} \max_{z \in S_{n+1}} |J_{n+1}^*(z) - \tilde{J}_{n_z}(z)| \\
&\leq \alpha^{\Delta t_{n+1}} \max_{z \in S_{n+1}} (|J_{n_z}^*(z) - \tilde{J}_{n_z}(z)| + \|J_{n+1}^* - J_{n_z}^*\|_\infty) \\
&\leq \max_{z \in S_{n+1}} \left( \alpha^{\Delta t_{n+1}} \max_{\{i_j\}_{j=0}^T \in W_{n_z}} A(\{i_j\}_{j=0}^T) + \alpha^{\Delta t_{n+1}} \|J_{n+1}^* - J_{n_z}^*\|_\infty \right) \\
&= \max_{\{i_j\}_{j=0}^T \in W_{n+1}} A(\{i_j\}_{j=0}^T).
\end{aligned}$$

The last equality is due to  $n+1 \leq m_{k+1} - 1$ , and  $\{m_j\}_{j=0}^k \subset \{\{i_j\}_{j=0}^T, n+1\} \subset \{0, 1, \dots, n+1\}$  for any  $\{i_j\}_{j=0}^T \in W_{n_z}$ . Therefore, Eq. 29 holds for all  $n \in (m_k, m_{k+1} - 1]$ . When  $n = m_{k+1} - 1$ , we also have the above relation for all  $z \in D_{n+1}$ :

$$\begin{aligned}
|\tilde{J}_{n+1}(z) - J_{n+1}^*(z)| &\leq \max_{z \in S_{n+1}} \left( \alpha^{\Delta t_{n+1}} \max_{\{i_j\}_{j=0}^T \in W_{n_z}} A(\{i_j\}_{j=0}^T) + \alpha^{\Delta t_{n+1}} \|J_{n+1}^* - J_{n_z}^*\|_\infty \right) \\
&= \max_{\{i_j\}_{j=0}^T \in W_{n+1}} A(\{i_j\}_{j=0}^T).
\end{aligned}$$

The last equality is due to  $n+1 = m_{k+1}$  and thus  $\{m_j\}_{j=0}^{k+1} \subset \{\{i_j\}_{j=0}^T, n+1\} \subset \{0, 1, \dots, n+1\}$  for any  $\{i_j\}_{j=0}^T \in W_{n_z}$ . Therefore, Eq. 29 also holds for  $n = m_{k+1}$  and this completes the induction.

We see that all  $\{i_j\}_{j=0}^T \in W_n$ , we have  $j \leq i_j \leq m_j$ , and thus  $j \leq S_{i_j} \leq j^{1/\theta}$ . By Lemma 10,

$$\lim_{n \rightarrow \infty} A(\{i_j\}_{j=0}^T \in W_n) = 0 \text{ w.p.1.}$$

Therefore,

$$\lim_{n \rightarrow \infty} \sup_{z \in D_n} |\tilde{J}_n(z) - J_n^*(z)| = 0 \text{ w.p.1.}$$

Since all states are updated infinitely often, and  $J_n^*$  converges uniformly to  $J^*$  with probability one, we conclude that:  $\lim_{n \rightarrow \infty} \|\tilde{J}_n - J_n^*\|_\infty = 0$  w.p.1. and  $\lim_{n \rightarrow \infty} \|\tilde{J}_n - J^*\|_\infty = 0$  w.p.1.

In both Steps S1 and S2, we have  $\lim_{n \rightarrow \infty} \|J_n - J_n^*\|_\infty = 0$  w.p.1<sup>5</sup>, therefore  $\mu_n$  converges to  $\mu_n^*$  pointwise w.p.1 as  $\mu_n$  and  $\mu_n^*$  are induced from Bellman updates based on  $J_n$  and  $J_n^*$  respectively. Hence, the sequence of policies  $\{\mu_n\}_{n=0}^\infty$  has each policy  $\mu_n$  as an  $\epsilon_n$ -optimal policy for the MDP  $\mathcal{M}_n$  such that  $\lim_{n \rightarrow \infty} \epsilon_n = 0$ . By Theorem 2, we conclude that

$$\lim_{n \rightarrow \infty} |J_{n, \mu_n}(z) - J^*(z)| = 0, \forall z \in S_n \text{ w.p.1.}$$

□

## E Proof of Theorem 8

We fix an initial starting state  $x(0) = z$ . In Theorem 7, starting from an initial state  $x(0) = z$ , we construct a sequence of Markov chains  $\{\xi_i^n; i \in \mathbb{N}\}_{n=1}^\infty$  under minimizing control sequences

<sup>5</sup>The tilde notion is dropped at this point.

$\{u_i^n; i \in \mathbb{N}\}_{n=1}^\infty$ . By convention, we denote the associated interpolated continuous time trajectories and control processes as  $\{\xi^n(t); t \in \mathbb{R}\}_{n=1}^\infty$  and  $\{u^n(t); t \in \mathbb{R}\}_{n=1}^\infty$  respectively. By Theorem 1,  $\{\xi^n(t); t \in \mathbb{R}\}_{n=1}^\infty$  converges in distribution to an optimal trajectory  $\{x^*(t); t \in \mathbb{R}\}$  under an optimal control process  $\{u^*(t); t \in \mathbb{R}\}$  with probability one. In other words,  $(\xi^n(\cdot), u^n(\cdot)) \xrightarrow{d} (x^*(\cdot), u^*(\cdot))$  w.p.1. We will show that this result can hold even when the Bellman equation is not solved exactly at each iteration.

In this theorem, we solve the Bellman equation (Eq. 9) by sampling uniformly in  $U$  to form a control set  $U_n$  such that  $\lim_{n \rightarrow \infty} |U_n| = \infty$ . Let us denote the resulting Markov chains and control sequences due to this modification as  $\{\bar{\xi}_i^n; i \in \mathbb{N}\}_{n=1}^\infty$  and  $\{\bar{u}_i^n; i \in \mathbb{N}\}_{n=1}^\infty$  with associated continuous time interpolations  $\{\bar{\xi}^n(t); t \in \mathbb{R}\}_{n=1}^\infty$  and  $\{\bar{u}^n(t); t \in \mathbb{R}\}_{n=1}^\infty$ . In this case, randomness is due to both state and control sampling. We will prove that there exists minimizing control sequences  $\{u_i^n; i \in \mathbb{N}\}_{n=1}^\infty$  and the induced sequence of Markov chains  $\{\xi_i^n; i \in \mathbb{N}\}_{n=1}^\infty$  in Theorem 7 such that

$$(\bar{\xi}^n(\cdot) - \xi^n(\cdot), \bar{u}^n(\cdot) - u^n(\cdot)) \xrightarrow{p} (0, 0), \quad (30)$$

where  $(0, 0)$  denotes a pair of zero processes. To prove Eq. 30, we first prove the following lemmas. In the following analysis, we assume that the Bellman update (Eq. 9) has minima in a neighborhood of positive Lebesgue measure. We also assume additional continuity of cost functions for discrete MDPs.

**Lemma 11** *Let us consider the sequence of approximating MDPs  $\{\mathcal{M}_n\}_{n=0}^\infty$ . For each  $n$  and a state  $z \in S_n$ , let  $v_n^*$  be an optimal control minimizing the Bellman update, which is referred to as an optimal control from  $z$ :*

$$\begin{aligned} v_n^* \in V_n^* &= \operatorname{argmin}_{v \in U} \{G_n(z, v) + \alpha^{\Delta t_n(z)} \mathbb{E}_{P_n} [J_{n-1}(y)|z, v]\}, \\ J_n(z, v_n^*) &= J_n^*(z) = G_n(z, v_n^*) + \alpha^{\Delta t_n(z)} \mathbb{E}_{P_n} [J_{n-1}(y)|z, v_n^*], \quad \forall v_n^* \in V_n^*. \end{aligned}$$

Let  $\bar{v}_n$  be the best control in a sampled control set  $U_n$  from  $z$ :

$$\begin{aligned} \bar{v}_n &= \operatorname{argmin}_{v \in U_n} \{G_n(z, v) + \alpha^{\Delta t_n(z)} \mathbb{E}_{P_n} [J_{n-1}(y)|z, v]\}, \\ J_n(z, \bar{v}_n) &= G_n(z, \bar{v}_n) + \alpha^{\Delta t_n(z)} \mathbb{E}_{P_n} [J_{n-1}(y)|z, \bar{v}_n]. \end{aligned}$$

Then, when  $\lim_{n \rightarrow \infty} |U_n| = \infty$ , we have  $|J_n(z, \bar{v}_n) - J_n^*(z)| \xrightarrow{p} 0$  as  $n$  approaches  $\infty$ , and there exists a sequence  $\{v_n^* \mid v_n^* \in V_n^*\}_{n=0}^\infty$  such that  $\|\bar{v}_n - v_n^*\|_2 \xrightarrow{p} 0$ .

**Proof** We assume that for any  $\epsilon > 0$ , the set  $A_\epsilon^n = \{v \in U \mid |J_n(z, v) - J_n^*(z)| \leq \epsilon\}$  has positive Lebesgue measure. That is,  $m(A_\epsilon^n) > 0$  for all  $\epsilon > 0$  where  $m$  is Lebesgue measure assigned to  $U$ . For any  $\epsilon > 0$ , we have:

$$\mathbb{P}(\{|J_n(z, \bar{v}_n) - J_n^*(z)| \geq \epsilon\}) = (1 - m(A_\epsilon^n)/m(U))^{|U_n|}.$$

Since  $1 - m(A_\epsilon^n)/m(U) \in [0, 1)$  and  $\lim_{n \rightarrow \infty} |U_n| = \infty$ , we infer that:

$$\lim_{n \rightarrow \infty} \mathbb{P}(\{|J_n(z, \bar{v}_n) - J_n^*(z)| \geq \epsilon\}) = 0.$$

Hence, we conclude that  $|J_n(z, \bar{v}_n) - J_n^*(z)| \xrightarrow{p} 0$  as  $n \rightarrow \infty$ . Under the mild assumption that  $J_n(z, v)$  is continuous on  $U$  for all  $z \in S_n$ , thus there exists a sequence  $\{v_n^* \mid v_n^* \in V_n^*\}_{n=0}^\infty$  such that  $\|\bar{v}_n - v_n^*\|_2 \xrightarrow{p} 0$  as  $n$  approaches  $\infty$ .  $\square$

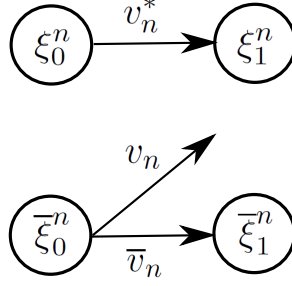


Figure 6: An illustration for Lemma 12. We have  $\bar{\xi}_0^n$  converges in probability to  $\xi_0^n$ . From  $\xi_0^n$ , the optimal control is  $v_n^*$  that results in the next random state  $\xi_1^n$ . From  $\bar{\xi}_0^n$ , the optimal control and the best sampled control are  $v_n$  and  $\bar{v}_n$  respectively. The next random state from  $\bar{\xi}_0^n$  due to the control  $\bar{v}_n$  is  $\bar{\xi}_1^n$ .

By Lemma 11, we conclude that  $\|J_n - J^*\|_\infty$  converges to 0 in probability. Thus,  $J_n$  returned from the iMDP algorithm when the Bellman update is solved via sampling converges uniformly to  $J^*$  in probability. We, however, claim that  $J_{n,\mu_n}$  still converges pointwise to  $J^*$  almost surely in the next discussion.

**Lemma 12** *With the notations in Lemma 11, consider two states  $\xi_0^n$  and  $\bar{\xi}_0^n$  such that  $\|\bar{\xi}_0^n - \xi_0^n\|_2 \xrightarrow{p} 0$  as  $n$  approaches  $\infty$ . Let  $\bar{\xi}_1^n$  be the next random state of  $\bar{\xi}_0^n$  under the best sampled control  $\bar{v}_n$  from  $\bar{\xi}_0^n$ . Then, there exists a sequence of optimal controls  $v_n^*$  from  $\xi_0^n$  such that  $\|\bar{v}_n - v_n^*\|_2 \xrightarrow{p} 0$  and  $\|\bar{\xi}_1^n - \xi_1^n\|_2 \xrightarrow{p} 0$  as  $n$  approaches  $\infty$ , where  $\xi_1^n$  is the next random state of  $\xi_0^n$  under the optimal control  $v_n^*$  from  $\xi_0^n$ .*

**Proof** We have  $\bar{v}_n$  as the best sampled control from  $\bar{\xi}_0^n$ . By Lemma 11, there exists a sequence of optimal controls  $v_n$  from  $\bar{\xi}_0^n$  such that  $\|\bar{v}_n - v_n\|_2 \xrightarrow{p} 0$ . We assume that the mapping from state space  $S_n$ , which is endowed with the usual Euclidean metric, to optimal controls in  $U$  is continuous. As  $\|\bar{\xi}_0^n - \xi_0^n\|_2 \xrightarrow{p} 0$ , there exists a sequence of optimal controls  $v_n^*$  from  $\xi_0^n$  such that  $\|v_n - v_n^*\|_2 \xrightarrow{p} 0$ . Now,  $\|\bar{v}_n - v_n\|_2 \xrightarrow{p} 0$  and  $\|v_n - v_n^*\|_2 \xrightarrow{p} 0$  lead to  $\|\bar{v}_n - v_n^*\|_2 \xrightarrow{p} 0$  as  $n \rightarrow \infty$ . Figure 6 illustrates how  $\bar{v}_n, v_n$ , and  $v_n^*$  relate  $\bar{\xi}_1^n$  and  $\xi_1^n$ .

Using the probability transition  $P_n$  of the MDP  $\mathcal{M}_n$  that is locally consistent with the original continuous system, we have:

$$\begin{aligned} \mathbb{E}[\xi_1^n \mid \xi_0^n, u_0^n = v_n^*] &= \xi_0^n + f(\xi_0^n, v_n^*)\Delta t_n(\xi_0^n) + o(\Delta t_n(\xi_0^n)), \\ \mathbb{E}[\bar{\xi}_1^n \mid \bar{\xi}_0^n, \bar{u}_0^n = \bar{v}_n] &= \bar{\xi}_0^n + f(\bar{\xi}_0^n, \bar{v}_n)\Delta t_n(\bar{\xi}_0^n) + o(\Delta t_n(\bar{\xi}_0^n)), \\ Cov[\xi_1^n \mid \xi_0^n, u_0^n = v_n^*] &= F(\xi_0^n, v_n^*)F(\xi_0^n, v_n^*)^T \Delta t_n(\xi_0^n) + o(\Delta t_n(\xi_0^n)), \\ Cov[\bar{\xi}_1^n \mid \bar{\xi}_0^n, \bar{u}_0^n = \bar{v}_n] &= F(\bar{\xi}_0^n, \bar{v}_n)F(\bar{\xi}_0^n, \bar{v}_n)^T \Delta t_n(\bar{\xi}_0^n) + o(\Delta t_n(\bar{\xi}_0^n)), \end{aligned}$$

where  $f(\cdot, \cdot)$  is the nominal dynamics, and  $F(\cdot, \cdot)F(\cdot, \cdot)^T$  is the diffusion of the original system that are assumed to be continuous almost everywhere. We note that  $\Delta t_n(\bar{\xi}_0^n) = \Delta t_n(\xi_0^n) = \gamma_t(\log(|S_n|)/|S_n|)^{\theta_{\varsigma\rho/d_x}}$  as  $\bar{\xi}_0^n$  and  $\xi_0^n$  are updated at the  $n^{th}$  iteration in this context, and the holding times converge to 0 as  $n$  approaches infinity. Therefore, when  $\|\bar{\xi}_0^n - \xi_0^n\|_2 \xrightarrow{p} 0$ ,  $\|\bar{v}_n - v_n^*\|_2 \xrightarrow{p} 0$ , we have:

$$\mathbb{E}[\bar{\xi}_1^n - \xi_1^n \mid \xi_0^n, \bar{\xi}_0^n, u_0^n = v_n^*, \bar{u}_0^n = \bar{v}_n] \xrightarrow{p} 0, \quad (31)$$

$$Cov(\bar{\xi}_1^n - \xi_1^n \mid \xi_0^n, \bar{\xi}_0^n, u_0^n = v_n^*, \bar{u}_0^n = \bar{v}_n) \xrightarrow{p} 0. \quad (32)$$



Since  $\bar{\xi}_1^n$  and  $\xi_1^n$  are bounded, the random vector  $\mathbb{E}[\bar{\xi}_1^n - \xi_1^n \mid \xi_0^n, \bar{\xi}_0^n, u_0^n = v_n^*, \bar{u}_0^n = \bar{v}_n]$  and random matrix  $Cov(\bar{\xi}_1^n - \xi_1^n \mid \xi_0^n, \bar{\xi}_0^n, u_0^n = v_n^*, \bar{u}_0^n = \bar{v}_n)$  are bounded. We recall that if  $Y_n \xrightarrow{p} 0$ , and hence  $Y_n \xrightarrow{d} 0$ , when  $Y_n$  is bounded for all  $n$ ,  $\lim_{n \rightarrow \infty} \mathbb{E}[Y_n] = 0$  and  $\lim_{n \rightarrow \infty} Cov(Y_n) = 0$ . Therefore, Eqs. 31-32 imply:

$$\lim_{n \rightarrow \infty} \mathbb{E}[\mathbb{E}[\bar{\xi}_1^n - \xi_1^n \mid \xi_0^n, \bar{\xi}_0^n, u_0^n = v_n^*, \bar{u}_0^n = \bar{v}_n]] = 0, \quad (33)$$

$$\lim_{n \rightarrow \infty} Cov(\mathbb{E}[\bar{\xi}_1^n - \xi_1^n \mid \xi_0^n, \bar{\xi}_0^n, u_0^n = v_n^*, \bar{u}_0^n = \bar{v}_n]) = 0, \quad (34)$$

$$\lim_{n \rightarrow \infty} \mathbb{E}[Cov(\bar{\xi}_1^n - \xi_1^n \mid \xi_0^n, \bar{\xi}_0^n, u_0^n = v_n^*, \bar{u}_0^n = \bar{v}_n)] = 0. \quad (35)$$

The above outer expectations and covariance are with respect to the randomness of states  $\xi_0^n, \bar{\xi}_0^n$  and sampled controls  $U_n$ . Using the iterated expectation law for Eq. 33, we obtain:

$$\lim_{n \rightarrow \infty} \mathbb{E}[\bar{\xi}_1^n - \xi_1^n] = 0.$$

Using the law of total covariance for Eqs. 34-35, we have:

$$\lim_{n \rightarrow \infty} Cov[\bar{\xi}_1^n - \xi_1^n] = 0.$$

Since

$$\mathbb{E}[\|\bar{\xi}_1^n - \xi_1^n\|_2^2] = \mathbb{E}[(\bar{\xi}_1^n - \xi_1^n)^T (\bar{\xi}_1^n - \xi_1^n)] = \|\mathbb{E}[\bar{\xi}_1^n - \xi_1^n]\|_2^2 + tr(Cov[\bar{\xi}_1^n - \xi_1^n]),$$

the above limits together imply:

$$\lim_{n \rightarrow \infty} \mathbb{E}[\|\bar{\xi}_1^n - \xi_1^n\|_2^2] = 0.$$

In other words,  $\bar{\xi}_1^n$  converges in 2<sup>th</sup>-mean to  $\xi_1^n$ , which leads to  $\|\bar{\xi}_1^n - \xi_1^n\|_2 \xrightarrow{p} 0$  as  $n$  approaches  $\infty$ .  $\square$

Returning to the proof of Eq. 30, we know that  $\xi_0^n = \bar{\xi}_0^n = z$  as the starting state. From any  $y \in S_n$ , an optimal control from  $y$  is denoted as  $v^*(y)$ , and the best sampled control from the same state  $y$  is denoted as  $\bar{v}(y)$ .

By Lemma 12, as  $\bar{u}_0^n = \bar{v}(\bar{\xi}_0^n)$ , there exists  $u_0^n = v^*(\xi_0^n)$  such that  $\|\bar{u}_0^n - u_0^n\|_2 \xrightarrow{p} 0$  and  $\|\bar{\xi}_1^n - \xi_1^n\|_2 \xrightarrow{p} 0$ . Let us assume that  $(\|\bar{u}_{k-1}^n - u_{k-1}^n\|_2, \|\bar{\xi}_k^n - \xi_k^n\|_2)$  converges in probability to  $(0, 0)$  upto index  $k$ . We have  $\bar{u}_k^n = \bar{v}(\bar{\xi}_k^n)$ . Using Lemma 12, there exists  $u_k^n = v^*(\xi_k^n)$  such that  $(\|\bar{u}_k^n - u_k^n\|_2, \|\bar{\xi}_{k+1}^n - \xi_{k+1}^n\|_2) \xrightarrow{p} (0, 0)$ . Thus, for any  $i \geq 1$ , we can construct a minimizing control  $u_i^n$  in Theorem 7 such that  $(\|\bar{\xi}_i^n - \xi_i^n\|_2, \|\bar{u}_i^n - u_i^n\|_2) \xrightarrow{p} (0, 0)$  as  $n \rightarrow \infty$ . Hence, Eq. 30 follows immediately:

$$(\bar{\xi}^n(\cdot) - \xi^n(\cdot), \bar{u}^n(\cdot) - u^n(\cdot)) \xrightarrow{p} (0, 0).$$

We have  $(\xi^n(\cdot), u^n(\cdot)) \xrightarrow{d} (x^*(\cdot), u^*(\cdot))$  w.p.1. Thus, by hierarchical convergence of random variables [30], we achieve

$$(\bar{\xi}^n(\cdot), \bar{u}^n(\cdot)) \xrightarrow{d} (x^*(\cdot), u^*(\cdot)) \text{ w.p.1.}$$

Therefore, for all  $z \in S_n$ :

$$\lim_{n \rightarrow \infty} |J_{n, \mu_n}(z) - J^*(z)| = 0 \text{ w.p.1.}$$

$\square$

## F Proof of Theorem 9

Fix  $n \in \mathbb{N}$ , for all  $z \in S$ , and  $y_n = \operatorname{argmin}_{z' \in S_n} \|z' - z\|_2$ , we have

$$\bar{\mu}_n(z) = \mu_n(y_n).$$

We assume that optimal policies of the original continuous problem are obtainable. By Theorems 7-8, we have:

$$\lim_{n \rightarrow \infty} |J_{n, \mu_n}(y_n) - J^*(y_n)| = 0 \text{ w.p.1.}$$

Thus,  $\mu_n(y_n)$  converges to  $\mu^*(y_n)$  almost surely where  $\mu^*$  is an optimal policy of the original continuous problem. Thus, for all  $\epsilon > 0$ , there exists  $N$  such that for all  $n > N$ :

$$\|\mu_n(y_n) - \mu^*(y_n)\|_2 \leq \frac{\epsilon}{2} \text{ w.p.1.}$$

Under the assumption that  $\mu^*$  is continuous at  $z$ , and due to  $\lim_{n \rightarrow \infty} y_n = z$  almost surely, we can choose  $N$  large enough such that for all  $n > N$ :

$$\|\mu^*(y_n) - \mu^*(z)\|_2 \leq \frac{\epsilon}{2} \text{ w.p.1.}$$

From the above inequalities:

$$\|\mu_n(y_n) - \mu^*(z)\|_2 \leq \|\mu_n(y_n) - \mu^*(y_n)\|_2 + \|\mu^*(y_n) - \mu^*(z)\|_2 \leq \epsilon, \forall n > N \text{ w.p.1.}$$

Therefore,

$$\lim_{n \rightarrow \infty} \|\bar{\mu}_n(z) - \mu^*(z)\|_2 = \lim_{n \rightarrow \infty} \|\mu_n(y_n) - \mu^*(z)\|_2 = 0 \text{ w.p.1.}$$

□